

UNIVERSITAT JAUME I  
Departament de Llenguatges i Sistemes Informàtics



Data selection and spectral-spatial characterisation  
for hyperspectral image segmentation.  
Applications to remote sensing.

Ph. D. dissertation  
Olga RAJADELL ROJAS

Supervisors  
Dr. Pedro GARCÍA SEVILLA  
and Prof. Filiberto PLA BAÑÓN

Castellón, December 2013



*To Rubén, Paqui and Fernando.*



## Acknowledgements

After five years of PhD, many different characters have joined me on the way. I want to make sure that the main characters get as much prominence as they deserve by keeping this section brief. For all those I will not name: thanks to everybody who gave me ideas, support, help and encouragement. It was all very necessary.

Nevertheless, some people deserve something longer. My supervisor professor Filiberto Pla who first endorsed me and trusted me to start this work. The role of the one that was my daily supervisor, Pedro García, has been just as important. He deserves a long sentence here for all the support, time and knowledge he offered, through Skype or in person. This work would have never started and succeeded without them. Thanks to all my colleagues from the computer vision group of UJI. M'agradaria dedicar unes paraules en especial a Yasmina i Mónica: us desitjo tot el millor. Adolfo, gràcies per orientar-me en més d'una ocasió. Gracias a Raúl que ha hecho posible que este manuscrito se vea algo más profesional. Investigadors com vosaltres no haurien de trobar-se en situació precària.

Thanks to Antonio Plaza for his interest, encouragement and disposition for discussion. I wish our country had many more researchers like you.

One word cannot express my gratitude to Robert Duin and Marco Loog who kindly welcomed me in their group. Jullie hadden een belangrijke rol in mijn promotie, die zonder jullie niet hetzelfde zou zijn. Ik ben jullie zowel professioneel als persoonlijk erg dankbaar. David, Coung, Alessandro, Yan, Laurence and Wan-Jui, bedankt. Special thanks to Veronika who, apart of being a friend, has helped me correcting this big manuscript.

Nothing would have been the same during the last year without working at the department of intelligent imaging at TNO (The Netherlands). Thanks specially to Jasper, Arvit, Auke, Coen, Gerard, Maarten, Willem (in alphabetical order) and Bernadetta for making the last year better.

I owe even more to those who always welcome me back. David, Ester e Israel. Muchas gracias por esperarme como si el tiempo no pasara. Thanks to Belén, for years and years of friendship.

Por último y más importante. A las recién llegadas Magda y Chloé, no saben lo que les queda por soportar. A los que sí lo saben, porque siempre han estado conmigo: Tora, Rubén, Paqui y Fernando. Mamá y Rubén, gracias por creer que podía hacerlo todo, a veces viene bien que alguien crea que eres el mejor, aunque no lo seas. Papá, has sido, sin lugar a duda, mi mejor profesor.

## Funding

El trabajo llevado a cabo ha sido financiado principalmente por la Fundacixa Castelló-Bancaixa mediante la beca FPI de la Fundació Bancaixa PREDOC/2007/20 y el proyecto P1-1B2007-48. Ha colaborado parcialmente el Ministerio de Ciencia e Innovación a través los proyectos CSD2007 00018 from Consolider Ingenio 2010, AYA2008-05965-C04-04 from Spanish CICYT and MTM2009-14500-C02-02.

Sobre todo, gracias a la Universidad Jaume I por haber decidido destacarse de entre otras apoyando constantemente a sus investigadores, en la medida de lo posible. Dada la tesitura en la que se en-

cuentra el país, no cabe más que desearle que encuentre la manera de mantenerse fiel a sus principios y que los que me suceden cuenten con el mismo apoyo que yo he tenido.

# Preface

## Abstract

Lately image analysis have aided many discoveries in research. This thesis focusses on the analysis of remote sensed images for aerial inspection. It tackles the problem of segmentation and classification according to land usage. In this field, the use of hyperspectral images has been the trend followed since the emergence of hyperspectral sensors. This type of images improves the performance of the task but raises some issues. Two of those issues are the dimensionality and the interaction with experts. We propose enhancements overcome them. Efficiency and economic reasons encouraged to start this work. The enhancements introduced in this work allow to tackle segmentation and classification of this type of images using less data, thus increasing the efficiency and enabling the design task specific sensors which are cheaper. Also, our enhancements allow to perform the same task with less expert collaboration which also decreases the costs and accelerates the process.

**Keywords:** hyperspectral, remote sensed images, classification, segmentation, characterization, texture.

## Resumen

El análisis de imágenes ha impulsado muchos descubrimientos en la ciencia actual. Esta tesis se centra en el análisis de imágenes remotas para inspección aérea, exactamente en el problema de segmentación y clasificación de acuerdo al uso del suelo. Desde el nacimiento de los sensores hiperespectrales su uso ha sido vital para esta tarea ya que facilitan y mejoran sustancialmente el resultado. Sin embargo el uso de imágenes hiperespectrales entraña, entre otros, problemas de dimensionalidad y de interacción con los expertos. Proponemos mejoras que ayuden a paliar estos inconvenientes y hagan el problema más eficiente.





# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Basic concepts on hyperspectral image classification . . . . .	1
1.1.1 Hyperspectral imaging . . . . .	1
1.1.2 Image classification and segmentation . . . . .	2
1.1.3 Supervised/unsupervised/semisupervised classification . . . . .	4
1.2 Objectives . . . . .	5
1.2.1 Thesis overview . . . . .	5
1.2.2 Contributions . . . . .	7
<b>2 Theoretical Background</b>	<b>9</b>
2.1 Spectral-Spatial characterization . . . . .	9
2.1.1 Textural methods . . . . .	10
2.2 Classification . . . . .	11
2.2.1 $k$ -Nearest Neighbour classifier . . . . .	14
2.2.2 Support Vector Machine Classifiers . . . . .	14
2.2.3 Clustering . . . . .	15
2.2.4 Active learning . . . . .	15
2.3 Curse of dimensionality . . . . .	16
2.3.1 Band Selection . . . . .	16
<b>3 Spectral-Spatial image characterization</b>	<b>19</b>
3.1 Characterization using textural features . . . . .	21
3.1.1 Gabor filters . . . . .	21
3.1.2 Wavelets filters . . . . .	22
3.1.3 Gabor versus Wavelets filters . . . . .	22

3.2	Spectral-Spatial characterization based on Gabor filters . . . . .	25
3.2.1	Gabor filters over individual planes . . . . .	25
3.2.2	Gabor filters over complex planes . . . . .	26
3.2.3	Opponent features . . . . .	28
3.2.4	Analysis of the complexity of the representations . . . . .	29
3.2.5	Dyadic vs. fixed width scales . . . . .	29
3.3	Experimental results . . . . .	31
3.3.1	Experimental setup . . . . .	31
3.3.2	Band selection issues . . . . .	32
3.3.3	Classification results . . . . .	34
3.3.4	Scale analysis . . . . .	34
3.3.5	Segmentation . . . . .	44
3.4	Conclusions . . . . .	56
<b>4</b>	<b>Training selection</b>	<b>57</b>
4.1	Training selection . . . . .	58
4.1.1	Mode seeking clustering . . . . .	60
4.1.2	Classification . . . . .	65
4.2	Experimental results . . . . .	69
4.2.1	Influence of the spatial information . . . . .	70
4.2.2	Classification by active learning . . . . .	71
4.2.3	Segmentation results . . . . .	75
4.3	Conclusions . . . . .	80
<b>5</b>	<b>Conclusions</b>	<b>85</b>
5.1	Future work . . . . .	86
<b>6</b>	<b>Sinopsis de la tesis</b>	<b>87</b>
6.1	Motivación . . . . .	87
6.2	Objetivos . . . . .	88
6.3	Contribuciones . . . . .	89
6.3.1	Esquema alternativo para la creación de mapas de superficie terrestre . . .	89
6.3.2	Reducción de la dimensionalidad mediante el análisis de la información . .	89
6.3.3	Selección de los datos de entrenamiento del sistema . . . . .	90
6.3.4	Difusión del trabajo de investigación . . . . .	90
6.4	Conclusiones . . . . .	90
6.5	Líneas de trabajo futuras . . . . .	91
	<b>Bibliography</b>	<b>93</b>
<b>A</b>	<b>Datasets</b>	<b>103</b>
<b>B</b>	<b>Publications</b>	<b>109</b>

# List of Figures

1.1	Decomposition of the spectrum of the light in its different wavelengths. . . . .	2
1.2	Representation of a point in a scene by its corresponding response to the light recorded by a hyperspectral sensor. . . . .	3
1.3	Pixel classification allows to assign a label to each pixel and that forms a map of labels which is a segmentation of the land cover represented by the original hyperspectral image. . . . .	4
1.4	Traditional classification-segmentation scheme (left) against the suggested in this thesis (right). . . . .	6
1.5	Thesis overview scheme: the objectives previously numbered are represented in circles and within the method proposed to carry them out. . . . .	7
3.1	Visualization of the filter bank with $M = 6$ and $N = 4$ in the spatial frequency domain. . . . .	23
3.2	Wavelet decomposition expressed in the spatial frequency domain for the two levels of analysis using the Daubechies-4 filters . . . . .	24
3.3	Visualization of the responses of one band of AVIRIS dataset to the Gabor filter bank with $M = 6$ and $N = 4$ . . . . .	27
3.4	Graphical scheme on how combinations of the planes are made for the complex plane method. . . . .	28
3.5	Graphical scheme on how combinations of the filtered responses are made for obtaining opponent features. . . . .	29
3.6	Number of features per method as the number of bands involved increase. . . . .	30
3.7	Gabor filter banks with $N = 4$ and (a) dyadic tessellation $M = 6$ , (b) constant width tessellation of $M = 8$ . . . . .	31
3.8	Classification rates for individual bands using Gabor filters. (a) AVIRIS (b) CHRIS-PROBA (c) ROSIS at the University of Pavia. The range of water absorption and the low SNR bands have been marked in grey. Selected bands using WALUMI with $B = 4$ are marked with red lines. . . . .	33

3.9	Pixel classification rates for different characterization methods over AVIRIS and CHRIS-PROBA databases. (a)(c) Dyadic tessellation. (b)(d) Constant tessellation.	35
3.10	For the AVIRIS dataset, pixel classification rates using independently features from the same scale of the filter bank. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	37
3.11	For the CHRIS-PROBA dataset, pixel classification rates using independently features from the same scale of the filter bank. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	38
3.12	For the AVIRIS dataset, pixel classification rates using features starting from the first scale independently, and progressively joining the following scales. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	39
3.13	For the CHRIS-PROBA dataset, pixel classification rates using features starting from the first scale independently, and progressively joining the following scales. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	40
3.14	For the AVIRIS dataset, pixel classification rates using features starting from the last scale independently, and progressively joining the following scales from highest to lowest. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	41
3.15	For the CHRIS-PROBA dataset, pixel classification rates using features starting from the last scale independently, and progressively joining the following scales from highest to lowest. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation. . . . .	42
3.16	Pixel classification rates for different characterization methods over AVIRIS and CHRIS-PROBA databases using a reduced number of features according to scale analysis performed. (a)(c) Dyadic tessellation. (b)(d) Constant tessellation. . . . .	43
3.17	Kappa coefficient against number of bands used for the different characterization methods over AVIRIS and CHRIS-PROBA databases using a number of bands $B \in [1..10]$ and (a) a complete dyadic filter and (b) only with $M = 1, 2, 3, 4$ . . . . .	45
3.18	Segmentation results for AVIRIS dataset using a complete filter bank and a different number of bands $B \in [2..4]$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above. . . . .	47

3.19	Segmentation results for AVIRIS dataset using a filter bank with $M = 1, 2, 3, 4$ and a different number of bands $B \in [2..4]$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above. . . . .	49
3.20	Segmentation results for AVIRIS dataset using $B = 4$ and a filter bank with (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above. . . . .	51
3.21	Segmentation results for CHRIS-PROBA dataset using a filter bank with $B = 4$ and (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6, 7$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above. . . . .	53
3.22	Segmentation results for ROSIS dataset using a filter bank with $B = 4$ and (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6, 7$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above. . . . .	55
4.1	Diagram representing the flow of the new classification scheme that includes the training selection. Each step is numbered as seen in the text. Notice that the selection can be performed with internal improvements or with post-process improvements. . . . .	59
4.2	Three first features for all classes in the AVIRIS database (a) when no spatial coordinates are included (b) when spatial coordinates are included as features. . . . .	62
4.3	Three first features for two classes in the AVIRIS database (a) when no coordinates are included (b) when spatial coordinates are included as features. . . . .	62
4.4	Representation of 66 clusters and their centers (white dots) over AVIRIS groundtruth (first row) and over the corresponding cluster result (second row). The clustering uses spatial coordinates in both cases. In (a-c) without modification and (b-d) overweighed. . . . .	64
4.5	(a) White dots are the 66 cluster centers resulting from clustering procedure; (b) the previous result after applying the spatial criterion, 58 pixels remain. Both represented as white points over AVIRIS database groundtruth. . . . .	64
4.6	(a) 66 Clusters represented as an image using random colors; (b) classification result when labels given for the modes are propagated to the rest of the cluster; (c) groundtruth with the modes found marked on white. . . . .	66
4.7	Training selection example for extension of the scheme to SVM necessities. Two modes are highlighted with a point, from the mode a line is drawn to the furthest sample within the cluster and a circle marks the distance in which samples will be selected to extend the label of the mode. . . . .	68

4.8	Learning curve of two small standard data sets to validate the extension of the scheme for SVM using label propagation. Classification is presented in terms of error rate versus the size of training data in number of samples selected by the scheme suggested. The error introduced in the label propagation is also encountered and the training data size stands for the number of samples labeled. . . . .	69
4.9	Learning curve of the $k$ -NN classifier in terms of error rate when increasing the size of training data in number of samples selected by the suggested scheme showing the impact of including the spatial coordinates as features. . . . .	70
4.10	Learning curve of classification in terms of error rate versus the size of training data in number of samples selected by the scheme suggested with the two improving alternatives compared with the usual random pick. Classification with semi-supervised clustering is also included. In all cases, features consist of 10 spectral features and when a classification is performed, $k$ -NN classifier with $k=1$ is used. The results are shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases. . . . .	72
4.11	Learning curve of classification in terms of error rate versus the size of training data in number of samples. Random pick and selection technique of the training data are used. For the two of them, after selecting the training in the same way, two different type of features are used for classification: 10 spectral features and spatial-spectral features. In both cases, the classification is performed using a $k$ -NN classifier with $k=1$ . This is shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases. . . . .	74
4.12	Learning curve of classification in terms of error rate versus the size of training data in number of samples selected by the scheme suggested. In all cases, features consist of 10 spectral features but the classification is performed using three different classification algorithms. $k$ -NN classifier with $k=1$ , SVM with label propagation and semi-supervised clustering. The results are shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases. . . . .	76
4.13	Representation of the 70 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The right column corresponds to the classification results for each case on the left respectively. The error, misclassified pixels, is represented in white. For AVIRIS dataset. . . . .	77
4.14	Representation of the 220 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The images on the right corresponds to the classification results for each case represented on the left image, misclassified pixels are represented in white. For CHRIS-PROBA dataset. . . . .	78
4.15	Representation of the 220 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The images on the right corresponds to the classification results for each case represented on the left image, misclassified pixels are represented in white. For HYMAP dataset. . . . .	79

---

A.1	Hyper-spectral image AVIRIS (92AV3C over the Indian Pines Test Site). (a) Color composition; (b) Ground-truth; (c) Target classes to be recognized. . . . .	104
A.2	Hyper-spectral image captured by the CHRIS-PROBA system. (a) Color composition; (b) Ground-truth; (c) Target classes. . . . .	105
A.3	Hyper-spectral image captured by the ROSIS system at the University of Pavia. (a) Color composition; (b) Ground-truth; (c) Target classes. . . . .	106
A.4	Hyper-spectral image captured by the HyMap imaging spectrometer. (a) Color composition; (b) Ground-truth; (c) Target classes. . . . .	107





# List of Tables

2.1	Chronological review of spatial methods for classification and segmentation of hyperspectral images. . . . .	13
3.1	Classification rates (in percentage) band number 4 of the AVIRIS database with two different textural features and the spectral features. . . . .	25
3.2	Accuracy per class for the 17 classes classification of the AVIRIS dataset using the complete filter bank and different number of bands. . . . .	48
3.3	Accuracy per class for the 17 classes classification of the AVIRIS dataset using a filter bank with $M = 1, 2, 3, 4$ and different number of bands $B \in [2..4]$ . . . . .	50
3.4	Accuracy and kappa per class for the 17 classes classification of the AVIRIS dataset using $B = 4$ and a filter bank with (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6$ . . . . .	52
3.5	Accuracy and kappa per class for the 10 classes classification of the CHRIS-PROBA dataset using $B = 4$ and a filter bank with (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6$ . . . . .	54
3.6	Accuracy and kappa per class for the 10 classes classification of the ROSIS dataset using $B = 4$ and a filter bank with (a) $M = 1, 2$ , (b) $M = 1, 2, 3, 4$ , (c) $M = 1, 2, 3, 4, 5, 6$ . . . . .	54
4.1	Comparison between the spatial strategies suggested in this chapter. . . . .	65
4.2	Properties of each data set used in the experiments: Name, Number of classes present in the data set (NC), Size of the data set in samples (S), Maximum number of samples per class (MaxS), Minimum number of samples per class (MinS) and Dimensionality (D) . . . . .	68
4.3	Accuracy per class for the 17 classes classification of the AVIRIS dataset using semi-supervised clustering classification on 12 features (ten spectral features and two spatial coordinates). For a training set of 0.3%, 2% and 4% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white. . . . .	81

---

4.4	Accuracy per class for the 10 classes of the CHRIS-PROBA dataset using semi-supervised clustering classification with 12 features (ten spectral features and two spatial coordinates). For a training set of 0.2%, 0.35% and 0.6% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white. . . . .	82
4.5	Accuracy per class for the 7 classes of the HYMAP dataset using semi-supervised clustering classification with 12 features (ten spectral features and two spatial coordinates). For a training set of 0.5%, 2.3% and 4.7% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white. . . . .	83
A.1	Selected bands using WaLuMi for AVIRIS, Chris-Proba, ROSIS and HyMap for $B$ varying from 1 to 10. Note that the selection is not incremental but most of bands are often repeated. . . . .	108

# Chapter 1

## Introduction

This chapter contains a short introduction to the main concepts which are necessary to understand the context of this thesis. The introduction to the basic concepts will be followed by a description of the thesis objectives.

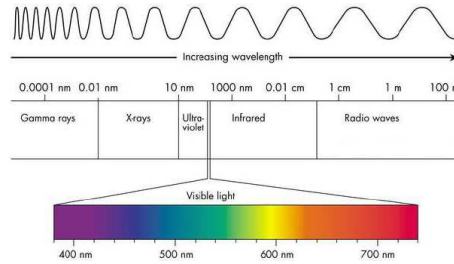
### 1.1 Basic concepts on hyperspectral image classification

#### 1.1.1 Hyperspectral imaging

A standard camera is a sensor that records the response of the scene to the visible part of the electromagnetic spectrum. The visible spectrum corresponds to the primary colours Red, Green and Blue. Therefore, the colour images obtained can be interpreted by the human eye. However, the spectrum of the light is wider than the visible range, see Figure 1.1 for a complete representation. When the light comes into contact with different mediums (water, rock, wood, corn fields) the response from all the spectrum is different. Despite the fact that the visible range responses equally for two media, for example soil and brown flooring, and the human eye sees the same color, the responses from the rest of the spectrum are different.

A hyperspectral sensor is able to capture a richer response of the scene to the visible and non-visible electromagnetic spectrum. The spectrum is a continuous signal and the sensor make it discrete by acquiring responses of ranges of consecutive wavelengths that cover the entire spectrum. The finer the ranges from which the responses are obtained, the higher the number of responses acquired. This is called spectral resolution of the sensor. The acquisition is made per point of the scene, which in an image is known as a pixel. When a sensor measurement is presented in matrix form, it obtains an image with three dimensions, two spatial dimensions that localized the point (pixel) in the scenario and a third spectral dimension that is composed by the responses to the light sensed by the sensor on that point. One independent acquisition, a pixel on an image, represents a point in the real scenario. The smaller real area that a pixel represents the more pixels are needed to cover the entire scenario but the resulting representation is more detailed. This property of the

sensor is the spatial resolution. Goetz et al. defined **Imaging spectrometry** as the simultaneous acquisition of images in many narrow, contiguous spectral bands [4].



**Figure 1.1:** Decomposition of the spectrum of the light in its different wavelengths.

The origin of the hyperspectral imaging concept is traced back to 1980, when A. F. H. Goetz and his colleagues at NASA's Jet Propulsion Laboratory developed the optical instrument called Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [86]. This sensor for Earth observation can record the visible and near-infrared spectrum of the reflected light resulting in more than 200 spectral bands.

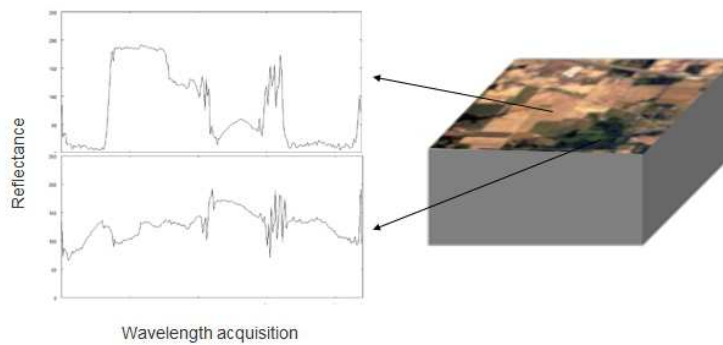
Examples of hyperspectral airborne imaging systems are AVIRIS [86], Hyperspectral Digital Imagery Collection Experiment (HYDICE) [85], Reflective Optics System Imaging Spectrometer (ROSIS) [85], Airborne Real-time Cueing Hyperspectral Enhanced Reconnaissance (ARCHER) [31] or HyMap [21]. One of the hyperspectral sensors currently operating in space are Hyperion (USA, 2000) [79]. The hyperspectral sensors typically cover a range of 0.4 to 2.5  $\mu\text{m}$  using 115 to 512 spectral channels, with a spatial resolution varying from 0.75 to 20 m/pixel for airborne sensors, and from 5 to 506 m/pixel for satellite sensors (see Figure 1.1).

Nowadays, hyperspectral imaging systems produce hundreds to thousands of spectral channels (bands). Because this type of images provides much richer and finer spectral information than traditional images, each pixel is represented by a spectral signature that characterizes uniquely the underlying surface. However, it also increases dramatically the volume of data. Hyperspectral data sets are often referred to as datacubes because of their 3-dimensional nature (two spatial and one spectral dimension). The pixel is the spatial unit and can be represented as a high-dimensional vector across the wavelength dimension containing the reflectance spectrum (see Figure 1.2). Since different substances exhibit different spectral signature, therefore, hyperspectral imaging is a well-suited technology for image classification.

### 1.1.2 Image classification and segmentation

**Hyperspectral image classification** can be defined as the identification of areas in a scene captured by a hyperspectral sensor [23]. It is an important task in many application domains such as:

- Agriculture: monitoring the development and health of crops.



**Figure 1.2:** Representation of a point in a scene by its corresponding response to the light recorded by a hyperspectral sensor.

- **Mineralogy:** identification of the composition of grounds for either localization or monitoring substances present on the surface.
- **Environment surveillance:** detecting changes on the environment is a task where hyperspectral imaging is helping to develop new techniques for climate change watching.
- **Military surveillance:** for detecting objects or substances that are hidden from the naked eye.

Towards these applications, different tasks need to be faced:

- **Dimensionality reduction:** reduction of the dimensionality of the input hyperspectral scene in order to facilitate following processing tasks.
- **Target or anomaly detection:** searching the pixels containing rare spectral signatures.
- **Classification/segmentation:** matching pixels/regions with a label to generate a land-cover map.
- **Spectral unmixing:** estimating the fraction of the pixel area covered by each material present in the scene.

Among those tasks this thesis tackles mainly the problem of classification/segmentation of hyperspectral images making use of dimensionality reduction techniques.

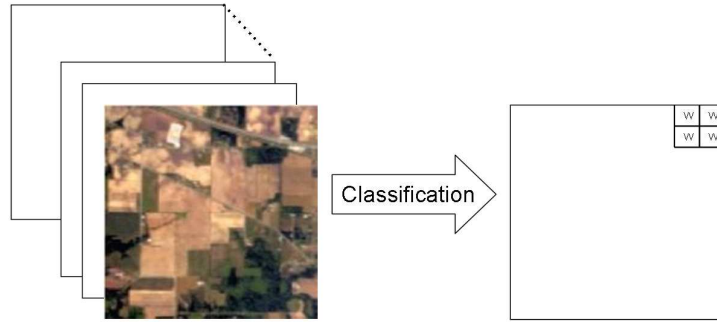
**Segmentation** is the division of the input image into non-overlapping regions, each of which is considered to be homogeneous with respect to some criterion of interest, often intensity or texture [92]. Hence, **Classification/segmentation** refers to the identification of the different class regions within the hyperspectral image. As mentioned in this introduction, the pixel is the spatial unit. That means the smaller unit for which the sensor records an independent spectral measure. This makes it also suitable to become the unit for classification and segmentation of the whole image.

**Pixel characterization.** In the context of image analysis, characterizing an image means to describe it using numerical features. This process is aimed at dealing with the content of images automatically [84]. As the measurements provided by the sensor are given per pixel, it is a straight forward idea to use pixels as the characterization unit.

Images have traditionally been characterized with the spectral signature of its pixels [43] [30] [1] [33]. An image  $I$  with spatial size  $N \times M$  and  $B$  spectral bands can be defined as the set:

$$I = \{\bar{x}_{ij} \in \mathbb{R}^B, i = 1, \dots, N, j = 1, \dots, M\}.$$

Each  $\bar{x}_{ij}$  is called feature vector as it contains the features describing a pixel in the image. Given a set of classes  $Y = \{y_1, y_2, \dots, y_C\}$ , **pixel classification and segmentation** consists of assigning each vector  $\bar{x}_{ij}$  a class  $y_c$ . The label (class) obtained for each pixel can be represented in the space of the scene allowing to obtain a segmentation and classification map of the target image [84] (see Figure 1.3). Although the feature vector can simply consist of the spectral signature of the pixel, it can also be extracted by techniques that use whether the surrounding pixels in the image itself or values obtained by filtering.



**Figure 1.3:** Pixel classification allows to assign a label to each pixel and that forms a map of labels which is a segmentation of the land cover represented by the original hyperspectral image.

### 1.1.3 Supervised/unsupervised/semisupervised classification

Formally, **classification** means assigning a class to an input [55]. Notice that classes receive a name or identifier, thus, the word class and label are often equally used in this context. Classification can be divided into unsupervised and supervised techniques.

The goal of **unsupervised** classification is to find a structure in the data  $X = \{x_1, \dots, x_n\}$ . Clustering is a typical unsupervised classification task. In this case the user has not foreknowledge of the classes existent within the data.

On the other side, **supervised** classification aims to find a mapping from a set of known labels  $Y = \{y_1, \dots, y_c\}$  to  $X$  by learning from observed data-label pairs [27]  $(x_i, y_j)$ ,  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, c\}$ .

This implies the assumption that the user knows the set  $Y$  of classes existent in the classification problem and that each of them can be described in the multidimensional space and this description is found by using data with a known class. This is called training data. Later unknown data is compared with the description and assigned a class according to the description that fits it better [52].

A new scenario has been introduced halfway between supervised and unsupervised classification, **semi-supervised classification**. In this last case only few information is given. In this scenario the foreknowledge is not enough for finding a reliable description of the classes and the number of classes present in the data,  $Y$ , may be also missing. However the information available can still provide a guideline [75].

## 1.2 Objectives

In literature there is a plentiful number of techniques for classification and segmentation of hyperspectral landscape images that make use of the entire spectral information of the datasets to perform pixel classification. The use of the entire spectral signature makes those methods difficult to scale as sensors quickly increase their spectral resolution. We propose to tackle this problem decreasing the dimensionality of the data and overcoming the lack of spectral information by adding spatial information.

In this field, techniques that use spatial information to improve the segmentation results are also found. However, they perform first pixel wise classification using the entire spectral signature of the pixels and apply afterwards corrections that make use of spatial information [106] [107] [64] [10]. We suggest a different approach: characterize pixels using spatial techniques and perform pixel wise classification on the characterization instead of on the spectral signature (Figure 1.4). Thus, the result of the classification is directly the final segmentation result.

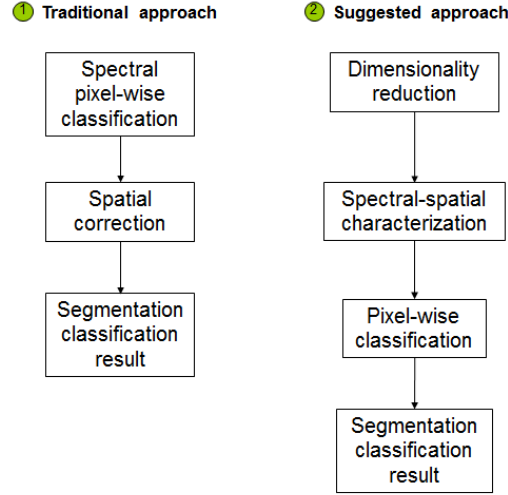
Finally, another common problem in the field of classification and segmentation of hyperspectral landscape images, is the lack of labeled data. Data needs to be labeled by an expert to train classifiers used in these techniques. The expert collaboration is always slower and more expensive than desired. We suggest a technique to reduce the training data so that the expert collaboration can be reduced.

Our objectives can be summarized in:

1. Improving the state of the art results of pixel classification.
2. Preventing the usage of large amounts of data to avoid dimensional problems.
3. Designing a technique to decrease the amount of labeled data needed while keeping the performance.

### 1.2.1 Thesis overview

Figure 1.5 represents an overview of the objectives of this thesis. The first objective of this work is to improve the state of the art results of pixel classification. We suggest including spatial information in the characterization of pixels to improve the pixel classification. However, hyperspectral



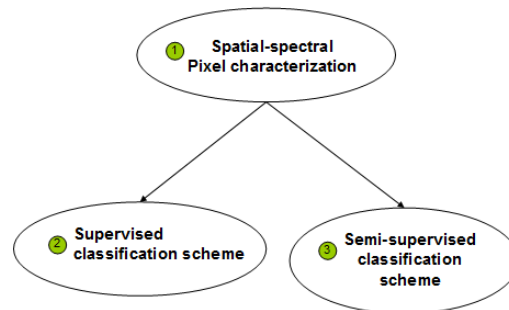
**Figure 1.4:** Traditional classification-segmentation scheme (left) against the suggested in this thesis (right).

data poses a dimensionality problem. Thus the second objective is developed by a scheme that allows to reduce the amount of data needed by introducing spatial information. This is not only a computational problem. Dealing with images with very high dimension implies expensive sensors and large times of data transmission too. Detecting in advance which information can be used in order to perform a specific task facilitates building specific sensors that are cheaper than the general ones. Besides, it also allows to transmit smaller amounts of data and eventually decreases the computation time.

Finally, the lack of labeled data is also a concern faced in this thesis. Labeled data is necessary for training systems but it is expensive and time consuming. Recall that in this specific case the expert involves a group of people and their equipment moving along a large area to determine the type of soil found in that location. Decreasing the amount of labeled data needed is of high interest and the third of our objectives. The approach suggested in this manuscript is a method to select training data according to an unsupervised analysis of the data. The selected data aims at representing a higher amount of data. Because of being more representative, the size of the training can decrease while achieving equal performances.

A background on the existing techniques is described in Chapter 2 where we introduce dimensionality reduction, Spectral-Spatial characterization and textural methods. In the same chapter the classification techniques are also reviewed. Our contributions on pixel characterization are summarized in Chapter 3 with the classification and segmentation results when using it on different datasets. Chapter 4 condenses the training selection technique, with different variants and the results on different hyperspectral datasets. At the end two appendices are included. Find first in Appendix A the description of the hyperspectral datasets used over the thesis experimentation. Ap-





**Figure 1.5:** Thesis overview scheme: the objectives previously numbered are represented in circles and within the method proposed to carry them out.

pendix B includes all the publications that support the contributions of this thesis and that have been condensed in Chapters 3 and 4 of this thesis.

### 1.2.2 Contributions

One of the objectives of this PhD was to obtain scientific impact by progressively divulging the progress of the research objectives. This was achieved by either conferences talks and journal publications. These publications can be found in the Appendix B, at the end of this thesis.

We presented a scheme for segmentation-classification of hyperspectral remote sensed images. This scheme includes data dimensionality reduction and spectral-spatial pixel characterization. It aims at improving land-use classification results decreasing significantly the number of spectral bands needed thanks to an adequate description of the spatial characteristics of the image. Requiring less data allows building task-specific sensors that decrease the costs. This idea, summarized in Chapter 3, was introduced in the following publications:

- O. Rajadell, P. Garcevilla and F. Pla., "Textural Features for Hyperspectral Pixel Classification". IbPRIA 2009, Lecture Notes in Computer Science 5524, pp.208-216.
- O. Rajadell, P. Garcevilla and F. Pla., "Scale Analysis of Several Filter Banks for Color Texture Classification". ISVC 2009, Lecture Notes on Computer Science 5876, pp.509-518.
- O. Rajadell, P. Garcevilla and F. Pla., "Filter Banks for Hyperspectral Pixel Classification of Satellite Images". CIARP 2009, Lecture Notes in Computer Science 5856, pp.1039-1046.
- O. Rajadell, P. Garcevilla and F. Pla., "On the Influence of Spatial Information for Hyperspectral Satellite Imaging Characterization". IbPRIA 2011, Lecture Notes in Computer Science 6669, pp.460-467.

- O. Rajadell, P. Garcevilla and F. Pla., "Spectral-Spatial Pixel characterization using gabor filters for hyperspectral image classification". *Geoscience and Remote Sensing Letters, IEEE*, vol.10, no.4, pp.860-864, July 2013.

Another objective was finding one solution to the concern of the expert collaboration. Our suggestion consisted of an unsupervised method for selecting training data. The addition of this method in the previous scheme achieves better classification and segmentation results while the necessary amount of labeled data is reduced. The method with its extensions is summarized in Chapter 4. However, the methodology was introduced, improved and extended in various publications:

- O. Rajadell, P. Garcia-Sevilla, V.C. Dinh and R.P.W. Duin, "Semi-supervised hyperspectral pixel classification using interactive labeling". 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) 2011, pp.1-4. Best paper award.
- O. Rajadell, P. Garcia-Sevilla, V.C. Dinh and R.P.W. Duin, "Selection of samples for active labeling in semi-supervised hyperspectral pixel classification". *Image and Signal Processing for Remote Sensing XVII* 2011, SPIE Proceedings Vol. 8180, pp.154-162.
- O. Rajadell, P. Garcia-Sevilla, "Training Selection with Label Propagation for Semi-supervised Land Classification and Segmentation of Satellite Images". *ICPRAM 2012, Lecture Notes in Computer Science* 204, pp.181-192.
- C.V. Dinh, M. Loog, R. Leitner, O. Rajadell, R.P.W. Duin, "Training data selection for cancer detection in multispectral endoscopy images". *ICPR 2012, Proceedings* pp.161-164.
- O. Rajadell, P. Garcia-Sevilla, V.C. Dinh and R.P.W. Duin, "Improving hyperspectral pixel classification with unsupervised training data selection". *Geoscience and Remote Sensing Letters, IEEE*, accepted, to be published.

# Theoretical Background

This thesis addresses the problem of classification and segmentation of remotely sensed hyperspectral images by using pixel classification. When pixels are the units of classification they can be characterized by a vector composed of their spectral signature sensed by the hyperspectral sensor. However, this vector can be enriched or replaced by a characterization that includes not only spectral information. Here, different methods of characterization are discussed. Furthermore, new classification methods are introduced. Finally, one of the main objectives is dealing with the dimensionality problem by improving the pixel characterization.

## 2.1 Spectral-Spatial characterization

Hyperspectral images contain richer spectral information than RGB images but the feature vectors extracted can still be enriched by adding information that is not present in the individual spectral signature. Each pixel represents a spatial point and the scene is composed of the set of all connected points. As sensors have evolved in acquiring finer spectral information, they have also improved their spatial resolution. While increasing the spectral information leads to acquiring a wider range of wavelengths, higher spatial information means that each pixel represents a smaller region of the scene. Hence, the detail of the representation increases and images exhibit spatial relations among neighboring pixels.

This information is very useful because a substance in a scene can be described by its spectral signature and its appearance. Recent trends in hyperspectral image classification detect the integration of spatial information in the data as a desired goal for improving the classification results [23][2][30]. Therefore, a wide range of techniques have arisen that study the integration of spatial and spectral information in the image analysis by combining or merging the spectral information with the information derived from properties of neighboring pixels within the feature vector in order to classify pixels according to both criteria.

There are several ways of including the neighbourhood information. Lately, morphological profiles [51][77] [10][68] and Markov fields [3][7][64] have been used very successfully. The definition

of a morphological segmentation method was first proposed by M. Pesaresi and J. A. Benediktsson in [69]. Mathematical morphology is a theory for the analysis of spatial structures in image data, it is based on pixels intensity and the idea is trying to characterize image structures by their morphological intrinsic characteristics. Extended Morphological Profiles (EMP) use opening and closing morphological transforms in order to isolate bright (opening) and dark (closing) structures in images, where bright/dark means brighter/darker than the surrounding pixels in the image. In this manner, each pixel characterization can include information about the size, shape and orientation of the structure it belongs to. This is obtained by using Structuring Elements (SE) with different sizes and shapes to model structures in the image [66]. However, the same type of objects may appear brighter than their neighbourhood in some parts of the image but darker in others. Analyzing the spatial information independently of their gray-level value is a desirable objective to pursue the characterization of the statistical relations between neighboring pixels. EMP cannot provide complete spatial information for an image scene [97].

Another widely used strategy in literature to integrate spatial information is Markov Random Fields (MRF), which model the statistical continuity among neighboring pixels. The intuitive idea behind this is that for a given pixel, its closest neighbours belong to the same class with a high probability, this is called class smoothness principle (CSP). MRF have been used within different classification procedures like Bayesian classification [53], Maximum A Posteriori framework using SVM [3] and lately MRF has been integrated with discriminative classifiers for computational efficiency [64]. MRF-based methods have proved to be a powerful tool for contextual image analysis. However, they traditionally require an iterative optimization step, which is time consuming.

It is important to mention that the described methods add to the spectral feature vector new extracted spatial features, and classify each pixel using all data. This leads to the increase of the dimension posing the so called Hughes effect [50] or curse of dimensionality.

Thus, it is of interest to find new methods to combine spatial and spectral information that overcome the mentioned disadvantages.

### 2.1.1 Textural methods

A joint spatial and spectral analysis has been identified as a desired goal and in that direction segmentation maps or descriptions of the neighbourhood of the pixels have been included into the classification in different ways. Nonetheless, when the spatial resolution increases the detail increases and areas exhibit patterns, that is, texture. This is an important property of the image that may help to recognize material according to its appearance.

**Texture** is the local statistical property of a region [81]. Texture classification is an active yet difficult topic in image processing, although there have been a lot of research efforts on it over the past several decades. The methods applied to texture classification have been widely discussed in literature. A review can be found in [81].

In [45] statistical methods based on Gaussian Markov random fields (GMRF) and Gibbs distribution models were proposed. Those models characterize the grey levels between nearest neighboring pixels by a statistical relationship. Although they yield good classification results, the algo-

rithms work by dividing the soil under investigation into small patches and feature extraction and classification are performed patch by patch.

Haralick et al proposed a method based on second-order texture statistics, the co-occurrence matrix [48]. This method was investigated by Tsai et al. [34] and Huang and Zhang [104] for including the spatial information in classification of hyperspectral data. In [34], texture images are generated using four measurements to describe the GLCM: Angular Second Moment, Contrast, Entropy and Homogeneity. Then, a Principal Component Analysis (PCA) is applied on the obtained texture images, and the Principal Components (PCs) are selected as features for Maximum Likelihood (ML) classification. Huang and Zhang [104] suggested performing Non-negative Matrix Factorization (NMF) feature extraction, followed by extracting spatial information using four measurements for the GLCM and applying SVM classification using spatial and spectral stacked features. The experimental results did not improve the pixelwise ones. This may be explained by the fact that the considered remote sensing images did not contain (or contained only a few) textured regions.

Methods such as wavelet analysis [71], Gabor filtering [39] or Local Binary Patterns [76] were developed for grey level or colour images. The extension of texture analysis methods for multi-channel images has been generally faced as a multi-dimensional extension of the mono-channel techniques [14][109][34]. It was Healey et al. [54] who made one of the first proposals for using spatial information across spectral bands using Gabor filters. Opponent features were first described for colour images [54] and lately extended to be used over multi-channel images [93]. Opponent features combine spatial information across spectral bands at different scales by combining the responses of the filters applied separately to each channel. Despite of being an innovative proposal, it was originally applied only to patches of stationary textures and global energies obtained were used to characterize and then classify the whole patches.

One of the main contributions of this thesis is the introduction of texture in the pixel wise classification task by replacing the spectral features by texture features.

## 2.2 Classification

As mentioned in the introduction, regarding the prior knowledge available, classification can be separated into supervised, unsupervised and semi-supervised.

Classification is the process of finding a mapping from data  $X = \{x_1, \dots, x_n\}$  to a set of known labels  $Y = \{y_1, \dots, y_m\}$  [55]. In the case of supervised classification [27] the mapping is learned from known data-label pairs  $(x_i, y_j)$ ,  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, m\}$ . Unsupervised classification can also be performed when neither the data, nor the labels  $Y$  are known and the process aims at discovering the structure in the data and the different classes that may be present. Last but not least, in semi-supervised classification only little information is given, a few  $(x_i, y_j)$  pairs, but the complete set of labels  $Y$  is unknown. Hence, some information can be used to find a mapping, but the whole structure of the data remains yet to be discovered [75].

Supervised classifiers learn from labeled examples and classify new unlabeled samples. For this, training data is necessary to build classifiers and it is generally recognized that the more data is available to train the more reliable our classifiers will be. Swain et al. [95] recommended a

minimum of  $N + 1$  samples, where  $N$  is the number of dimensions of the data, that is, the size of the pixel feature vector. However, they established a practical increase to  $10N$  training samples per class to get statistically stable results.

Notice that labeled data is limited and often expensive to obtain [101]. Therefore, in semi-supervised scenarios, active learning is a widely used technique. Active learning techniques are used to iteratively enlarge a very reduced set of examples. In each step of this process the learning function selects more objects regarding a criterium and considering the previous step and asks the expert to label them enriching the classification process of the next step. The iterations run until some convergence criterion is achieved [75]. This methodology provides good results for those cases where the expert is willing to collaborate [98][65].

Unsupervised classification is a totally different paradigm. Nothing is known about the data, neither examples nor the classes present and no supervision is available. In this case, the approaches are radically different since the aim is to find a structure that can give an idea of how many groups (classes) can be found in the data and which are the samples included in them. One of the most used techniques is clustering.

From the point of view of the classification technique used to find the mapping between samples  $X$  and classes  $Y$ , there are two types of classification techniques: generative and discriminative.

Generative models are those that learn the conditional density functions  $p(x|y)$ , the probability that sample  $x$  has the label  $y$ , separately from the training data and make their predictions using the Bayes rules to calculate  $p(y|x)$ . Examples of such algorithms include Linear Discriminant Analysis (LDA) and Naive Bayes classifier. On the other side, discriminative models learn the posterior  $p(y|x)$  directly from the data, this means that the class-conditional densities are not explicitly modeled. This property of discriminative models is one of the reasons for using them rather than generative models [100]. Linear and logistic discrimination,  $k$ -Nearest Neighbours ( $k$ -NN), tree classifiers, neural networks, Support Vector Machines (SVM) and other kernel methods are discriminative learning models.

One of the most popular non-linear discriminative algorithms is the  $k$ -Nearest Neighbour classifier ( $k$ -NN), it is intuitive and easy to implement. The nearest neighbour decision rule assigns to an unclassified sample the label of the nearest set of previously classified samples. This rule is independent of the underlying joint distribution on the sample points and their classifications [96]. Recall that the goal of designing a recognition system is to classify future test samples which are likely to differ from the training samples used to build the classifier. Therefore, building a system that perfectly predicts the class of training samples is unlikely to perform well on new patterns. This situation is known as overfitting. With larger training data sets,  $k$ -NN classifier tends to overfit.

Support Vector Machine (SVM) is one the most successful non-linear discriminative classifiers in the remotely sensed hyperspectral image community. It is characterized by its ability to effectively deal with large input spaces using limited training samples. Moreover, classification does not involve any assumptions about data distribution. Its superiority over standard classifiers (statistical and neural) has been studied in [42]. SVM attempts to separate training samples belonging to different classes by tracing maximum margin hyperplanes in the space where the samples are mapped.

Year	Reference	Classifier	Spatial Method
2001	Pessaresi [69]	NN	MP
2002	Jackson et al. [53]	Bayes	MRF
2003	Mercier et al. [44]	SVM	Kernels
2004	Dell'Acqua et al [32]	Combined	MP
2005	Benediktsson [10]	NN	MP
2005	Farag et al [3]	SVM	MRF
2006	Camps-Valls et al [43]	SVM	Kernels
2007	Fauvel [67]	Combined	Kernels
2009	Tarabalka et al. [106]	SVM	Segmentation map
2010	Tarabalka et al. [107]	SVM	Segmentation map
2011	Jun Li et al. [64]	Bayes Based MLR	MRF MLL

**Table 2.1:** Chronological review of spatial methods for classification and segmentation of hyperspectral images.

In 2003, Mercier et al. suggested in [44] a modified kernel that took into consideration the spectral similarity between support vectors outperforming SVM based on classification of the hyperspectral data cube. Later, in [43], Camps et al. presented a framework based on composite kernel machines for enhanced classification of hyperspectral images which exploited the properties of Mercer's kernels to construct a family of composite kernels that easily combine spatial and spectral information. A review on SVM methods can be found in [30]. These novel SVM formulations represent significant developments in which spatial and spectral information can be easily integrated and analyzed by using proper kernel functions. However, the integration of spatial and spectral information is generally done through the combination of dedicated kernels to spectral and contextual information [43].

Another interesting new approach is combining the result of a classification with a previously calculated segmentation map. Each segmented region defines a spatial neighbourhood for all the pixels within this region. Having a few pixels reliably classified, the region they belong to can be classified too. Tarabalka et al. in [106] and [107] explore this idea towards the objective of getting more homogenous regions from a SVM pixelwise classification.

In the first case [106], the results of a pixel-wise SVM classification and a segmentation map obtained by partitional clustering are combined. This is achieved by performing a majority voting on the pixelwise spectral classification using adaptive neighbourhoods defined by the segmentation map. ISODATA and Gaussian mixture techniques for image segmentation were tested. Although there was still some segmentation noise in the classification map, it was reduced by a fixed-window-based postfiltering. In [107] the watershed segmentation algorithm is used instead to define the spatial structures that are used as adaptive neighbourhoods for context classification.

In Table 2.1 a summary of mentioned outstanding methods can be found.

### 2.2.1 $k$ -Nearest Neighbour classifier

This supervised classifier is particularly simple in concept. It assumes that pixels close to each other in feature space are likely to belong to the same class. In its simplest form an unknown pixel is labeled by examining the available training pixels in multi-spectral space and choosing the class most represented among a pre-specified number of nearest neighbours. The comparison essentially requires the distances from the unknown pixel to all training pixels to be computed. Suppose there are  $k_i$  neighbours labeled as class  $\omega_i$  out of  $k$  nearest neighbours for a pixel vector  $x$ , noting that  $\sum_{i=1}^M k_i = k$  where  $M$  is the total number of classes. In the basic kNN rule we define the discriminant function for the  $i^{th}$  class as  $g_i(x) = k_i$  and the decision rule is:  $x \in \omega_i$ , if  $g_i(x) > g_j \forall j \neq i$

### 2.2.2 Support Vector Machine Classifiers

The Support Vector Machine (SVM) supervised classification concept was introduced to remote sensing image classification by [47] and multiple reviews give helpful details [12][49][73]. To explain the usage of this classifier, let us consider a two class problem in a  $D$ -dimensional space with  $N$  samples. For all samples  $x_i \in \mathbb{R}^D$  the set of their corresponding labels are available,  $(x_i, y_i), i \in [1, N]$ . The SVM algorithm tries to find the hyperplane  $H_p$  that maximizes the margin to the closest training data points of both classes:

$$w \cdot x + b = 0 \forall x \in H_p \quad (2.1)$$

$$y_i(w \cdot x_i + b) > 1 \forall x_i \notin H_p, i \in [1, N] \quad (2.2)$$

Where  $w \in \mathbb{R}^D$  is the vector normal to the hyperplane and  $b \in \mathbb{R}$  the bias and the margin between  $x$  and  $H_p$ ,  $\forall x_i \notin H_p$ , is given by:

$$f(x) = \frac{|w \cdot x + b|}{\|w\|} \quad (2.3)$$

The vector  $w$  should be such that satisfies Eq.(2.2). The optimal hyperplane  $H_p$  is the one that maximizes  $\frac{2}{\|w\|}$ . This is equivalent to minimizing  $\frac{\|w\|}{2}$ . Hence it is a quadratic optimization problem:

$$\min \left[ \frac{\|w\|^2}{2} \right], \text{ fulfilling (2.2)} \quad (2.4)$$

As non-linearly separable data is very common, slack variables  $\xi$  are introduced to deal with misclassified samples and Eq.(2.2) becomes:

$$y_i(w \cdot x_i + b) > 1 - \xi_i \forall i \in [1, N], \xi_i \geq 0 \quad (2.5)$$

And the optimization problem:

$$\min \left[ \frac{\|w\|^2}{2} + C \sum_{i=1}^N \xi_i \right], \text{ fulfilling Eq.(2.5)} \quad (2.6)$$



Where the constant  $C$  is a regularization parameter that controls the amount of penalty.

For the classification of an unknown sample  $x_u$  it must be computed:  $y_u = \text{sgn}(w \cdot x_u + b)$ , where  $(w, b)$  are the hyperplane parameters found during the training process. Because the vectors in the optimization and decision rule equations always appear in pairs related by a scalar product, these products can be replaced by nonlinear functions of the pairs of vectors. Thus, the pixel vectors can be projected into a higher dimensional space  $\mathbb{H}$  where the linear separability of data may be improved:

$$\mathbb{R}^D \rightarrow \mathbb{H}x \rightarrow \Phi(x)x_i \cdot x_j \rightarrow \Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j) \quad (2.7)$$

Where  $\Phi$  is the non linear function that projects the feature vectors in the new space and  $K$  a kernel. The usage of  $K$  decreases the computational complexity by helping to avoid the computation of the scalar products. The kernel  $K$  should fulfill the Mercer condition [12]. The two most used kernels in hyperspectral image classification are the polynomial function and the Gaussian Radial Basis Function (RBF)[19].

### 2.2.3 Clustering

Clustering is an unsupervised classification technique. A cluster is a group of similar objects that are close in the feature space [83] and clustering is a technique used to group the objects according to certain criteria. To perform this task different algorithms are known. In general these algorithms can be divided into parametric and non-parametric [55]. Parametric techniques make assumptions about the shape of the clusters to find the structure within the analyzed data and divide it in several groups with the chosen shape.

Non-parametric methods make no model assumptions, that is, clusters have an arbitrary shape and the connectivity within groups of objects is found by hierarchical or density based procedures [103]. The first ones either aggregate or divide the data set according to some agreed measure. The latter considers the probability density function of the feature space and search for local maxima. Based on the local structure of the feature space, a number of samples are associated to the maxima found. Non-parametrical analysis is known to be more flexible and suitable when absolutely no information is known about the data to analyze [87] [55].

### 2.2.4 Active learning

Consider that, for any supervised learning system to perform well, it must often be trained on big amounts of labeled samples. These labels are sometimes very difficult, time-consuming, or expensive to obtain. Active learning is based on the idea that if the data provided to the learning algorithm is chosen according to its necessities, less training data would be necessary to perform equally well [91].

The goal of active learning is mainly to reduce the cost of acquiring large labeled training sets. In active learning methods the classifier selects new training examples that would maximize the learning about the unlabeled dataset in order to improve the classification accuracy. Active learning

is meant for those scenarios where there are sufficient unlabeled samples but it is expensive to obtain class labels [75].

This strategy has become the last trend in hyperspectral imaging classification [80] [25] [26] [99] [36]. However, all these approaches are applied in research on previously acquired data for which groundtruth is known. It is very important to recall that the application of these strategies in the classification of real unknown hyperspectral landscape imaging would involve a group of experts with the corresponding equipment moving in a vast land extension.

## 2.3 Curse of dimensionality

The impact on the classification of increasing the dimensions of the data has been studied in literature. It is known as the **curse of dimensionality or Hughes phenomenon** [50]. When the dimensionality increases, the volume of the space increases so the available data becomes sparse. Hence, with a fixed number of training samples, the predictive power reduces as the dimensionality increases. Therefore, to obtain a statistically reliable result, the amount of data needed grows significantly if the dimensionality of the data increases [9].

Multi-spectral data has many advantages. Thanks to the detailed analysis of the response to the light, different materia can be better described and then classified. Despite the high advantage in classification, as the spectral resolution increases, the correlation among the spectral data increases too. High correlated data includes redundant information. The less correlated the data is the more separable it is in the multi-dimension space where it has to be classified.

Therefore, hyperspectral imaging is high dimensional and that poses the curse of dimensionality. For facing higher dimensional problems more training data is necessary. This requirement presents a challenge in practice where data is not always available in the desired amount. Thus, keeping the number of features used as low as possible is important when reliable results are expected and only affordable numbers of training pixels are available.

### 2.3.1 Band Selection

There are two ways of reducing the size of the data: performing feature extraction or feature selection. The first reduce the dimensions by transforming the features into an alternative set of features by applying a transformation, as a result a new set of features is obtained in a different space with a smaller dimension [90] [58] [102] [61]. Feature selection does not transform the features but chose among them the ones that maximize certain criteria. Hence, the reduced set of features is a subset of the original one [72] [17] [89] [94]. Since the features here are the bands of the image, these methods are also known as band selection algorithms. Notice that reducing the dimension of the data decreases the correlation and redundancy but does not alter the original data. Sensors with higher spectral resolution are more expensive and the transmission of all data needs a longer time. The purpose for reducing the dimensionality is to be able to use a cheaper sensor that captures and transfer only the required information. Consequently we did not consider feature extraction methods, like PCA, because they need all data to be available.

## WaLuMI

WaLuMI band selection method will be used in the thesis to reduce the dimensionality of the datasets. This is an unsupervised band selection method introduced by Martinez-Usó et al. [72]. The WaLuMI method (Ward's Linkage using Mutual Information) is based on hierarchical clustering and groups bands with two criteria: minimizing the intra-cluster variance while maximizing the inter-cluster variance. The algorithm consists of defining a dissimilarity space among image bands, where a distance criterion is defined based on the mutual information between any pair of bands. Then a hierarchical clustering is performed in the defined dissimilarity space. In order to progressively construct a hierarchical family of derived clusters the method uses a linkage strategy with an inter-cluster distance as the objective function to optimize. Finally, a band representing each final cluster is chosen. The  $B$  selected bands from the final  $B$  clusters provide an adequate representation reducing redundancy as much as possible. Note that this is not an incremental process, that is, the bands selected for a given value  $B$  are not always included in the selection for  $B + 1$ .

Band selection was chosen over feature extraction because it preserves the original data. Hyperspectral sensors are expensive. On the other hand, systems are usually designed for performing a repetitive task. We want to show that for a given dataset, composed by certain classes, reducing the number of spectral bands is possible and the segmentation classification task can be performed without losing precision. Once proved, a task driven sensor can be built for the classification and segmentation of the classes present in the dataset. This sensor would be cheaper than the original and the task would be faster to perform. This is not possible if feature extraction is used since to repeat the procedure all spectral bands are needed to extract the features. On top of that, we chose WaLuMI because it is unsupervised. Because another objective is to reduce the collaboration of the expert, the fact that the band selection can be autonomously performed is a very interesting requirement. These two factors are necessary for accomplishing our objectives and WaLuMI fulfills both.



# Chapter 3

## Spectral-Spatial image characterization

We can find in literature different approaches to deal with the classification of hyperspectral images. One of the most frequently found is done by extracting spatial portions (patches) from the image and classifying the rest of the patches according to the patches given for training. In this case the unit of classification (sample) is the patch [54][76][56][105][15]. Nevertheless, the interest lies not only in classifying blocks (patches) of the image but also every pixel in a way that the result is a segmentation and classification of the image. For such a task pixels should be the unit of classification, then, the final result is a map of pixels with the corresponding class which is also a segmentation. Hence, the segmented parts are groups of spatially connected pixels with the same class.

Characterizing pixels means to describe them with a numerical vector. These vectors are the samples for classification. A pixel is a spatial location, the simplest way of characterizing a pixel is describing it with the spectral measurement taken by the sensor in that location. However, the increase of the spectral resolution of sensors poses a dimensionality problem. Thus, as the spectral resolution of sensors increases, pixels are described by a larger set of more correlated features. This is undesirable.

When the spatial resolution of multi-spectral images was not high enough, researchers used the entire feature vector and the efforts were focused at the classification stage. These types of processing often used neural networks [46][105], decision trees [110], Bayesian estimation [15][7] and kernel-based methods [42][35] for pixel-wise classification. In particular, Support Vector Machines (SVM) proved to obtain good performances in this task [30]. Since nowadays the spatial resolution has also improved, a joint spectral-spatial analysis was detected as a goal [63]. This technique is called spectral-spatial characterization and aims at obtaining one feature vector for each pixel in the image based on the spectral measurements and/or a series of values extracted from spatial operations involving neighboring pixels (spatial information). Nowadays, a wide range of techniques is known to include spatial information into the image characterization, such as morphological profiles [68] or Markov fields [65]. However, these methods fall into over-segmentation [97]. Over-segmentation is found when many small disconnected areas appear in the result such that the segmentation does

not consist of smooth areas. This can also be given in the form of salt and pepper noise, that is, isolated misclassified pixels.

Recently, several proposals have been made to face the over-segmentation problem with very good results. Tarabalka et al. [106] presented a spectral-spatial classification scheme that consists of a pixel-wise classification and a partitional clustering by a majority vote with adaptive neighborhoods. The result is a segmentation map that needs a spatial post regularization to reduce noise. This provides more homogeneous regions than a simple pixel-wise classification process but it is not yet suitable for images containing small classes since they may be missed. The same problem is observed in [107] where an extension of the watershed segmentation algorithm for hyperspectral images was presented in order to define the spatial structures. To deal with the segmentation of small regions, the same authors suggested in [108] to select the most reliable pixels from a pixel wise classification as markers to be used in a Minimum Spanning Forest Grown. This obtains a spectral-spatial classification map refined afterwards by majority voting within the spatially connected regions. However, all these methods do not yet tackle the problem of the increasing dimensionality and make use of the entire feature vector form with all the spectral values. Furthermore, in some cases, this vector is enriched with more information which increases the already large number of features per pixel.

This thesis aims to tackle the dimensionality problem by suggesting a change in the scheme of current methods (classification and post-processing). This work suggests combining the selection of bands with the spatial characterization. This is possible due to the increase of spatial resolution in images. The pipeline suggested includes reducing the dimension of the images selecting the most informative bands, characterizing pixels using spectral-spatial information and classifying them.

The main consequence of the increase of the spatial resolution is that, in an image, one pixel represents a smaller portion of the real scenario so the surface is represented detailed enough for appreciating texture. Texture is a pattern in the distribution of the pixels within a connected area [81]. Textural analysis has been widely discussed in the literature to study the spatial relationships in an image. Therefore, texture is now a convenient property that can introduce a better description of the surroundings of each pixel.

There exists a huge variety of methods [81] namely: co-occurrence matrices [48], wavelet analysis [71], Gabor filtering [39] or Local Binary Patterns [76]. They all were developed mainly for grey level images. The extension of texture analysis methods for multi-channel images has been generally faced as a multi-dimensional extension of the mono-channel techniques. Healey et al. [54] made one of the first proposals on how to use spatial information across spectral bands using Gabor filters. Opponent features were first described for color images [54] and lately extended to be used over multi-channel images [93]. Opponent features combine spatial information across spectral bands at different scales by combining the responses of the filters applied separately to each channel. Despite of being an innovative proposal, it was originally applied only to patches of stationary textures and global energies were used to characterize and then classify the whole patches. Recently, similar experiments have been carried out using a three-dimensional Gabor filter bank [8] improving past results. However, again, samples are taken as patches from the image.

### 3.1 Characterization using textural features

One way of characterizing image pixels is using numerical features extracted by techniques that substitute the values contained in the image (spectral measurements) by those obtained applying any sort of filtering or transformation that allows to describe any helpful characteristic of the pixels surroundings.

Let  $I^i(x, y)$  be the  $i^{th}$  band of an image containing  $B$  bands. Let  $f_k(x, y)$  be the  $k^{th}$  filter in a filter bank  $F$ . The response of a pixel  $(x, y)$  to a filter when it is applied to an image band is given by the convolution of the image band with the filter:

$$h_k^i(x, y) = I^i(x, y) * f_k(x, y), \text{ where } f_k(x, y) \in F = \{f_k(x, y)\}_{k=1}^K \quad (3.1)$$

Applying the entire filter bank  $F$ , we obtain a feature vector which is composed of the responses for that pixel to all filters in the filter bank, that is:

$$\psi_{x,y} = \{h_k^i(x, y)\}_{k=1, i=1}^{K,B} \quad (3.2)$$

Therefore each pixel can be described with a series of numbers that are representative of a the response of the location to a bank of filters. If the filter bank is aimed at analyzing texture, this value includes a textural descriptor for each pixel. There have been statistical approaches in literature to approach characterization of image texture (see Section 2.1.1). These methods obtain their features from the image grey levels (spectral features) and they very often assume stationarity in the texture in a wide-sense whereas the statistics are in most images only locally stationary. Consequently multiresolution approaches became more popular. A multiresolution analysis decomposes the texture across several scales and examines the texture at different resolutions. The advantage of the multiresolution analysis is that the signal is decomposed over many scales. A fusion of the features extracted at each scale gives a more robust description of the signal than features extracted at only one scale. Two well known multiresolution textural analysis textural methods are the Gabor filters and the Wavelet transform.

#### 3.1.1 Gabor filters

Gabor filters received attention in literature for being able to analyze both orientation and spatial frequency [81]. Thus, a Gabor filter bank is a set of two-dimensional Gabor filters, see Eq. (3.3). Each Gabor filter is characterized by a preferred orientation  $n$  and a preferred spatial frequency range  $m$  (here it will be referred as scale). The filter acts as a local band-pass filter with optimal joint localization properties in the image spatial domain and the spatial frequency domain [39]. They consist essentially of sine and cosine functions modulated by a Gaussian envelope that achieve optimal joint localization.

$$F_G = \{f_{m,n}(x, y)\}_{m=1, n=1}^{M,N} = \{f_k(x, y)\}_{k=1}^K \text{ and } K = M \times N \quad (3.3)$$

They can be defined by Eq. (3.4) and (3.5) where  $m$  is the index for the scale,  $n$  the orientation and  $u_m$  is the central spatial frequency of the scale [56].

$$f_{mn}^{real}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2+y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_mx \cos \theta_n + u_my \sin \theta_n)) \quad (3.4)$$

$$f_{mn}^{imag}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2+y^2}{2\sigma_m^2}\right\} \times \sin(2\pi(u_mx \cos \theta_n + u_my \sin \theta_n)) \quad (3.5)$$

In Figure 3.1 a Gabor filter bank is shown in the spatial frequency domain, this is exactly the visualization of a Gabor filter bank with  $M = 6$  and  $N = 4$ . Notice that each spatial frequency scale is analyzed in four different orientations, which correspond to each row of the figure.

We use three methods based on textural analysis using Gabor filters to characterize pixels. None of them had been previously used for pixel characterization, although two of them have been used for patch classification in multispectral images. As mentioned, hyperspectral images have improved both the spatial and spectral resolution. One target is to exploit the spatial relations between pixels to better characterize them as a consequence of the increasing spatial resolution. Note that same reasoning can be applied for the spectral domain. The relations between spectral bands should be also considered. While for the first target textural characterization methods will be used, for the second, the concept of inter-channel information introduced by Healey et al [54] will be included in two of the three characterization methods.

### 3.1.2 Wavelets filters

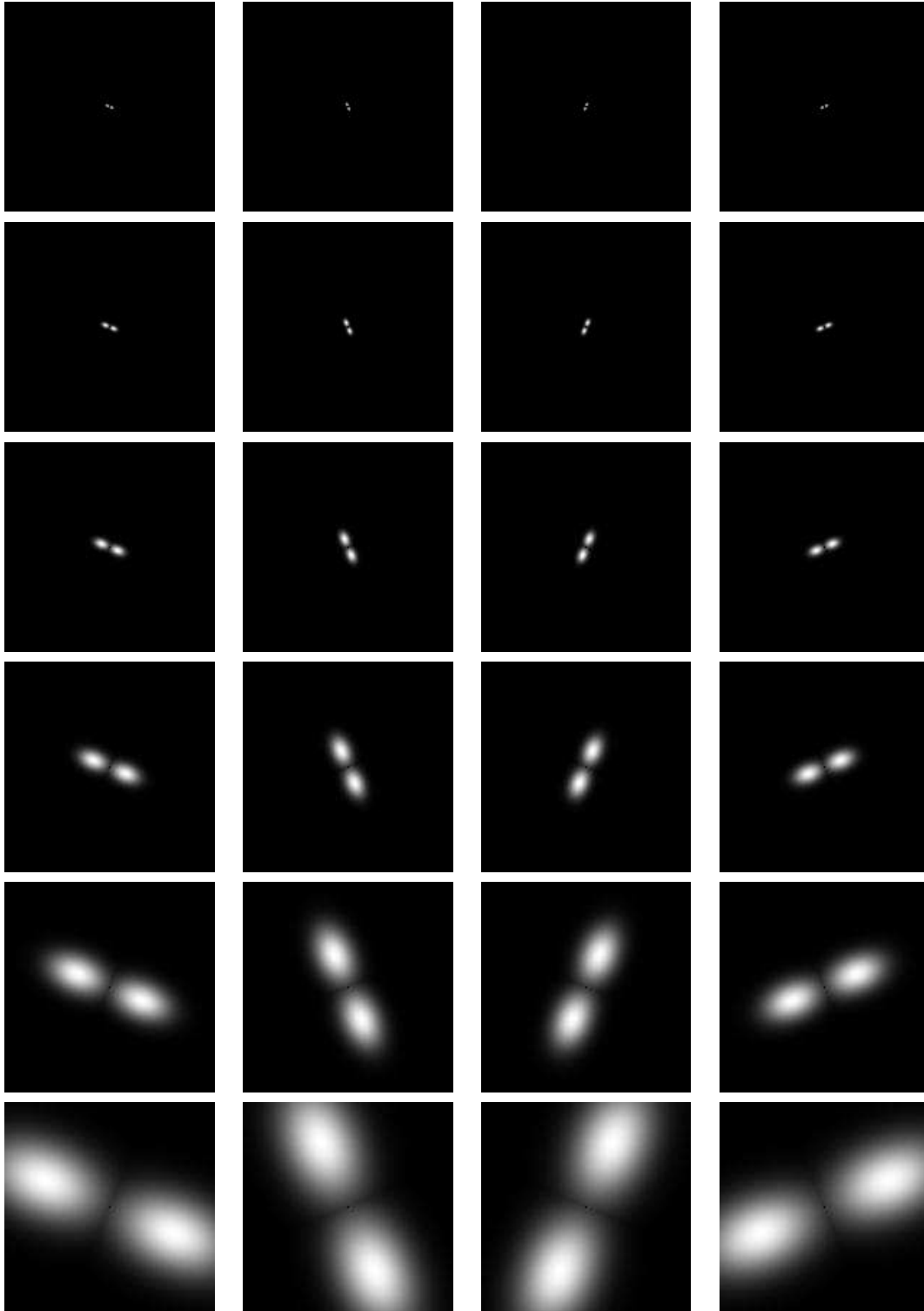
A wavelet transform can be specified by high-pass and low-pass filtering a signal using a particular wavelet filter. A class of compact wavelets functions which are nonzero over a finite range was discovered by Daubechies [24], and includes functions which range from being highly localized to being highly smoothed. The simplest case involves only four coefficients, see Figure 3.2. Daubechies wavelets can be formed with an even number of coefficients that satisfy certain orthogonality conditions and approximation conditions of order  $p$ . The wavelet transform has the property of giving both spatial frequency and image spatial information about an image.

### 3.1.3 Gabor versus Wavelets filters

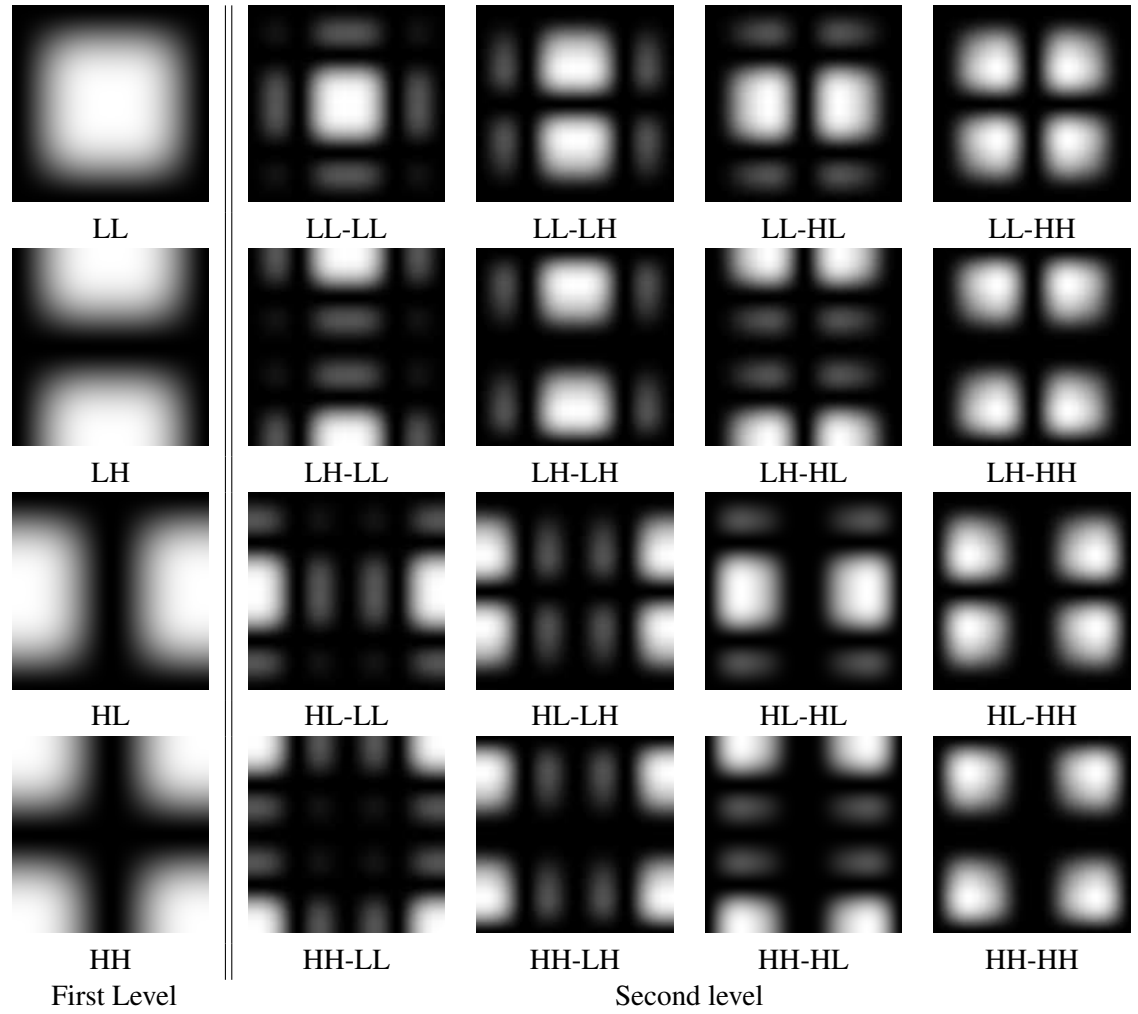
In order to test and compare the different textural features, we carried out a basic experiment over the AVIRIS dataset. The experiment consist of extracting features from only one band using Gabor and wavelets filter banks, Eq.( 3.2). The selected band is given by the band selection method, see Appendix A. The Gabor filter bank has four orientations and six scales and the wavelet decomposition is the the Daubechies-4 filters until three levels of decomposition. The methods provide a total of 24 and 84 features per pixel respectively for Gabor and Wavelets filter banks.

All samples (pixels) were divided into twenty independent sets keeping the prior probability of each class and the  $k$ -NN classifier was used to classify all sets taken in pairs, one used as the training





**Figure 3.1:** Visualization of the filter bank with  $M = 6$  and  $N = 4$  in the spatial frequency domain.



**Figure 3.2:** Wavelet decomposition expressed in the spatial frequency domain for the two levels of analysis using the Daubechies-4 filters

Features	Classification rate
Only spectral features	18.85 %
Wavelet features	27.77 %
Gabor features	41.58 %

**Table 3.1:** Classification rates (in percentage) band number 4 of the AVIRIS database with two different textural features and the spectral features.

set and the other as the test set [88]. Therefore, ten classification folds were performed without data dependencies among the folds. The mean rate of all the folds is reported in in Table 3.1.

Results in Table 3.1 show that both textural features outperform the single usage of spectral features. Wavelet features performed worse than expected. For each band it calculates 84 features but still the percentage of correct classification is just a bit better than the spectral ones. However, there is a significant increase in performance when using features calculated with Gabor filters. For further comparisons one can also consult [70][16][82]. Both literature and our small experiment suggest the same conclusion. Hence, for the rest of this thesis the efforts are focused on Gabor filter banks methods.

## 3.2 Spectral-Spatial characterization based on Gabor filters

In Section 3.1.1 we described Gabor filters as known in the literature. Here we suggest three different ways of using them for pixel characterization, none of them had been used before for this purpose.

### 3.2.1 Gabor filters over individual planes

This method filters each band of the image individually using the filter bank and characterizes each pixel with the response to it. Applying a filter bank, as seen in Eq. (3.1), is equivalent to pointwise multiplication ( $\bullet$ ) in the spatial frequency domain, Eq. (3.6). Notice that the spatial frequency decomposition of the image is obtained here using a Fourier Transform. We define the Fourier Transform of a function  $g$  as  $\hat{g}$ .

$$\hat{h}_k^i(x, y) = \hat{I}^i(x, y) \bullet \hat{f}_k(x, y), \text{ where } f_k(x, y) \in F = \{f_k(x, y)\}_{k=1}^K \quad (3.6)$$

Notice that Eq. (3.6) is defined in the spatial frequency domain. To obtain  $h_k^i(x, y)$  one must simply apply the inverse of the Fourier Transform. Then, the set of all the responses to the filter bank  $F$  form the feature vector for each pixel:

$$\psi_{x,y} = \{h_{mn}^i(x, y)\}_{\forall i,m,n} \quad (3.7)$$

where  $F$  is a Gabor filter bank,  $F_G$ , as defined in Eq. (3.3). They are used over images whose numbers are real. The Fourier transform of a real image is symmetrical. Thus, the filters applied should be symmetrical which means as well that the equivalent filters in the space domain are also real. Consequently the filters used are defined by the Eq. (3.4) since the imaginary part would be zero.

If  $M$  stands for the number of scales,  $N$  for the number of orientations and  $B$  for the number of channels that compose the image, the size of the pixel characterization vector is given by:

$$size(\psi_{x,y}) = MNB \quad (3.8)$$

For the analysis of an image, this image is decomposed into spatial frequencies applying the Fourier transform. Then, the filter bank is applied over the transform obtaining a feature vector for each pixel, Eq. (3.7). This way, the response of each pixel in the spatial domain represents the decomposition of the pixel spectral signature into what the value contributes to each scale (spatial frequency range) at a certain spatial orientation. Hence, filtering with this type of filter bank provides information for each pixel about the involvement of the pixel in different spatial frequencies what is a texture description.

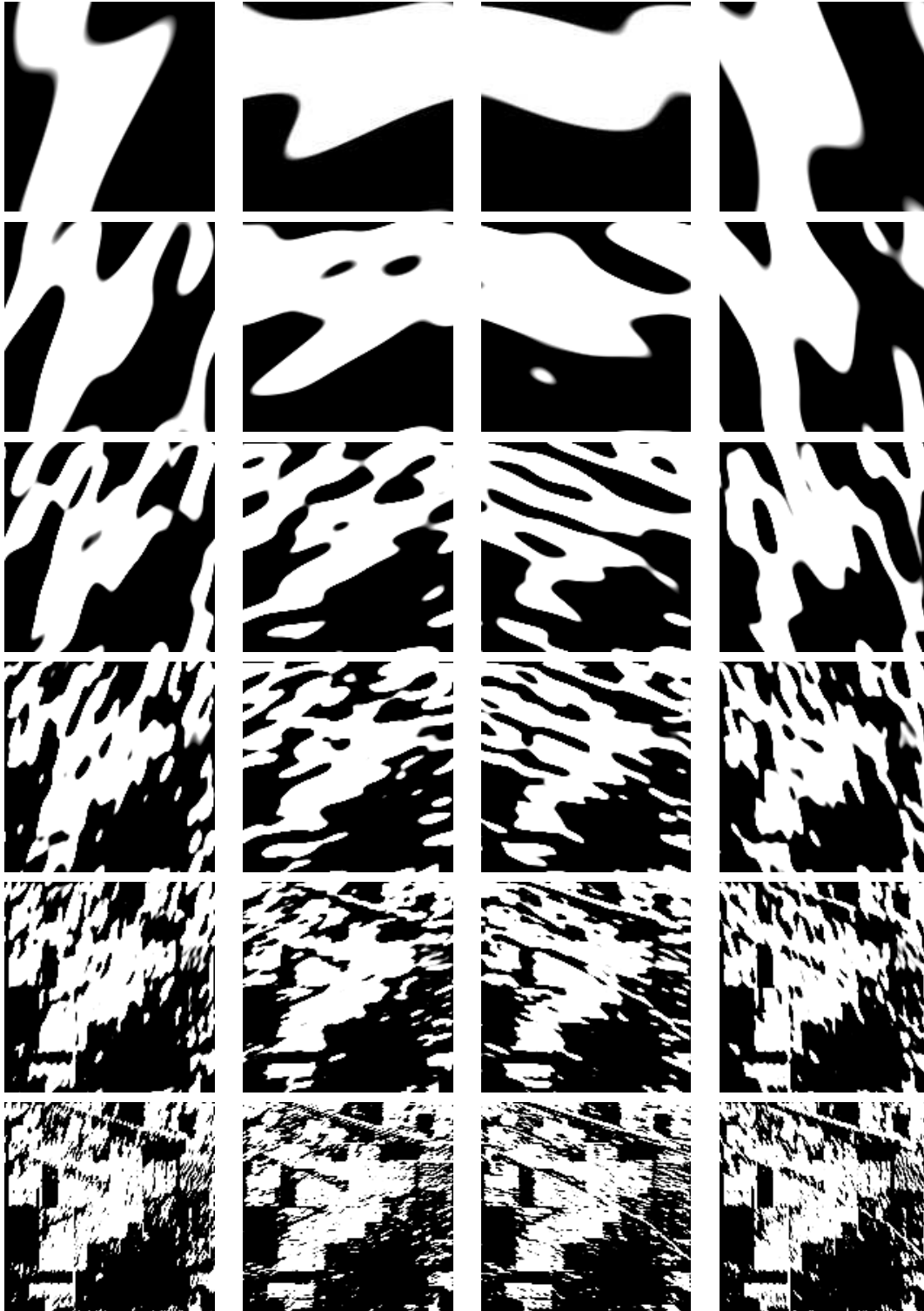
Consider the AVIRIS dataset (see Appendix A). According to the method description, the Fourier transform of each spectral band has to be multiplied by the bank filtered shown in Figure 3.1. This generates a series of responses that can be again brought to the spatial domain and visualized. In Figure 3.3 the responses for each pixel of one band of AVIRIS dataset for each of the corresponding filters in Figure 3.1 can be observed. Notice that intuitively the result of the analysis with these filters can be seen as an analysis of neighborhoods at different spatial frequency scales where the neighborhoods are set by the spatial frequency and orientation decomposition.

### 3.2.2 Gabor filters over complex planes

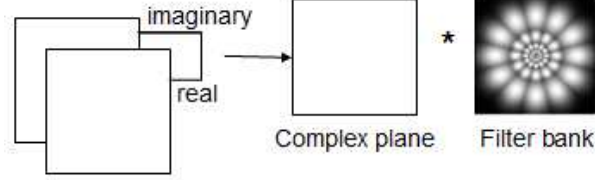
Recall that in the previous method each band (individual plane) is filtered by the filter bank to produce the outputs. Note that, in this way, Gabor filters are used as a direct extension of mono-channel images. Consequently this method fails to capture the inter-channel properties of a multi-channel image. Hence, to account for the inter-channel information, complex bands are proposed here.

A complex band is formed by merging two real bands into one complex band, one band as the real part and the other band as the imaginary part, see Figure 3.4. In this way, inter-channel information is involved within the characterization process. Since the channels to be analyzed are no longer real, their corresponding Discrete Fourier Transform (DFT) is not symmetrical. In this case, we suggest the usage of complex filters (non-symmetrical filters in the spatial frequency domain) as defined in Eq. (3.4) and (3.5). As a result, the number of orientations  $N$  is twice the number used in the previous method. Besides, all possible complex bands for the set of spectral bands are considered. Note that here not all spectral bands are considered but the reduced set selected by the band selection method.

The Gabor filter bank is applied over a complex plane as shown in Eq. (3.9), where  $I_i(x, y)$  and  $I_j(x, y)$  are the  $i^{th}$  and  $j^{th}$  spectral bands respectively and  $I_c^{ij} = I^i + I^j \mathbf{i}$  with  $\mathbf{i} = \sqrt{-1}$ .



**Figure 3.3:** Visualization of the responses of one band of AVIRIS dataset to the Gabor filter bank with  $M = 6$  and  $N = 4$ .



**Figure 3.4:** Graphical scheme on how combinations of the planes are made for the complex plane method.

$$\hat{h}_{mn}^{ij}(x, y) = \hat{I}_c^{ij}(x, y) \bullet \hat{f}_{m,n}(x, y) \quad (3.9)$$

In this case, the feature vector  $\phi_{x,y}$  for each pixel in the image is composed of the responses of all filters in the filter bank.

$$\phi_{x,y} = \{h_{mn}^{ij}(x, y)\}_{\forall i,j/i>j, \forall m,n} \quad (3.10)$$

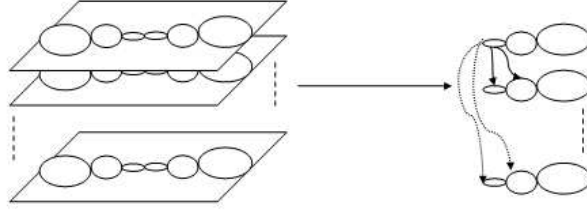
The size of the feature vector varies with the number of different complex bands. For each complex band one feature is obtained for each filter applied. Note also that, in this case, the response is a complex number and both the real and imaginary parts are used to characterize the image. This means that there are as many complex features as filters for each complex band and as many complex bands as combinations may be done with the  $B$  bands.

$$size(\phi_{x,y}) = 4MN \binom{B}{2} \quad (3.11)$$

### 3.2.3 Opponent features

The last method considered was suggested by Healey et al. [54] for classification of texture patches. Opponent features combine spatial information across spectral bands at different scales and are related to processes in human vision. They are computed from Gabor filters as the difference of outputs of two different filters. In other words, channels are first individually filtered and their responses are combined afterwards to obtain the opponent feature. The combination among responses is made for all pair of spectral bands  $i, j$  with  $i > j$  and  $0 \leq (m - m') \leq 1$  [93] as pictured on Figure 3.5 and according to the Eq.(3.12):

$$d_{mm'n}^{ij}(x, y) = h_{imn}^i(x, y) - h_{m'n}^j(x, y) \quad \forall i, j/i > j \quad \forall m, m', n/0 \leq (m - m') \leq 1 \quad (3.12)$$



**Figure 3.5:** Graphical scheme on how combinations of the filtered responses are made for obtaining opponent features.

Because we are interested in pixel classification, the feature vector  $\varphi_{x,y}$  for an individual pixel is obtained as the set of all opponent features for all pairs of spectral bands calculated for this pixel:

$$\varphi_{x,y} = \{d_{mm'n}^{ij}(x, y)\}_{\forall i,j/i>j, \forall m,m',n/0 \leq (m-m') \leq 1} \quad (3.13)$$

Hence, the size of the opponent feature vector depends on the number of orientations and scales and also takes into account pairs of bands:

$$\text{size}(\varphi_{x,y}) = (3M - 2)N \binom{B}{2} \quad (3.14)$$

Because a Gabor filter bank is used again to individually filter the channels that compose the image, filters are symmetrical in the spatial frequency domain, that is, real in the space domain. However, in this method the responses of the filters are combined later to introduce inter-channel information to the characterization process.

The relation of the size of the feature vector of the two methods discussed that use inter channel information is:

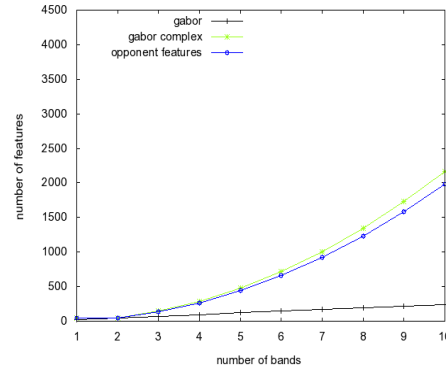
$$\text{size}(\phi_{x,y}) = \text{size}(\varphi_{x,y}) + (M + 2)N \binom{B}{2} \quad (3.15)$$

### 3.2.4 Analysis of the complexity of the representations

The methods suggested in the previous sections lead to a different number of features per pixel. In Figure 3.6 the number of features obtained is represented for a number of bands  $B \in [1...10]$ . This is important because of the so called Hughes phenomenon [50], the increase in the number of dimensions does not necessary lead to an improvement. Notice that whereas Gabor over individual planes (Gabor) increases the size linearly, the other two have an exponential factor which makes their dimensionality grow faster when the number of bands increase.

### 3.2.5 Dyadic vs. fixed width scales

Up to this point, Gabor filter banks have been used as defined in literature. The divisions of the spatial frequencies, called scales, are dyadic. This means that each scale is double the size of the



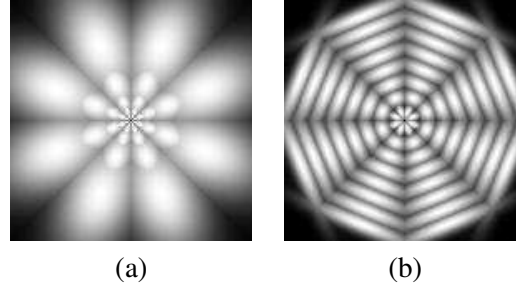
**Figure 3.6:** Number of features per method as the number of bands involved increase.

previous, see Figure 3.1. It is known that lower frequencies correspond to larger spatial regions and as the spatial frequency increase the representation of those frequencies in the image space domain correspond to smaller areas, until the highest spatial frequencies which correspond mostly to the noise contained in the image. Consequently, whereas lower spatial frequencies need to be analyzed with detail, higher ones do not. Notice that because we call scale to a division of the spatial frequency, when we refer to lower scale we mean low spatial frequencies and vice versa.

With the purpose of studying the contribution of the spatial frequency scales it is possible to re-define Gabor filters with constant width, that is, an equal partition of the different frequencies. For all characterization methods two different filter banks are taken into account: one with dyadic scales and one with constant width scales. When using a constant width tessellation of the spatial frequency domain, the width was chosen in order to have eight scales. Notice that the number of scales in a dyadic filter bank is given by the spatial size of the dataset, for the datasets considered here the dyadic filter varies the number of scales between six and seven. Regarding the number of orientations, it is always set to four. This means that we have four filters per scale when the filters are symmetrical, and eight filters per scale when the filters are non-symmetrical. It is important to note that certain degree of overlapping is introduced among the Gaussian distributions of the filters. In fact, Gaussian distributions are designed to overlap each other when achieving a value of 0.5. Both the number of orientation and the degree of overlapping are based on recommendations given in [11].

Figure 3.7 represents the summation of a complete Gabor filter bank in the spatial frequency domain for both cases, using dyadic and constant width scales. Observe that while the dyadic tessellation thoroughly analyzes low frequencies, given less importance to medium and higher frequencies, the constant one ensure an equal analysis of the entire spatial frequency domain.





**Figure 3.7:** Gabor filter banks with  $N = 4$  and (a) dyadic tessellation  $M = 6$ , (b) constant width tessellation of  $M = 8$ .

### 3.3 Experimental results

To validate each characterization method, the features obtained per pixel Eq.(3.7)(3.13)(3.10) are used for classification. Recall that a band selection method was used to obtain a reduced data set. On the different sets a characterization method is applied and a classifier is trained and tested. The classification results are shown as performance curves (overall accuracy) against the size of the set of bands. The set of bands is always the one selected by the selection method and showed in Figure A.1(c)-A.3(c). This representation will help to compare the performance between methods but also the gain obtained by increasing the size of the dataset. Each dataset has a number of known classes and one heterogeneous background class. For the classification experiments only the known classes are considered. In further experiments, for segmentation results, the heterogeneous background class will be also considered in order to obtain a segmentation map for the whole image. Nevertheless, per class accuracy is studied in those experiments and the classification statistics are given both considering and dismissing this special class.

#### 3.3.1 Experimental setup

We propose to change the typical classification-segmentation scheme (spectral classification, spatial correction, segmentation) to a setting where an unsupervised band selection method and pixel characterization precede the pixel-wise classification and the classification result is already the segmentation result (Figure 1.4). In a nutshell our scheme proceeds as follows:

1. Unsupervised band selection reduces the size of the dataset.
2. Characterization obtains a feature vector per pixel that replaces the spectral features of the traditional approach.
3. pixel-wise classification provides the final classification-segmentation result.

Two different settings of classification are used in the manuscript:

1. The labeled pixels in each image database are divided in twenty non-overlapping sets keeping the prior probability of each class. In this way, no redundancies are introduced and each set is a representative set of the bigger original one. Then, ten classification folds are carried out for each experiment and the mean of the error rates of these folds is taken as the final performance of the classifier. Each classification fold uses one of the twenty sets for training and another as test set. Therefore, each set is never used twice in the same experiment. This methodology was already used in [72] and [88] in order to increase the statistical independence among the classification folds.
2. In a different setting the set of samples is randomly divided into two sets. One set with 5% of the data is used as the training input for the classifier and the other, with the rest of the data, as test set to evaluate. This a very common setting used in [30] [62] [97].

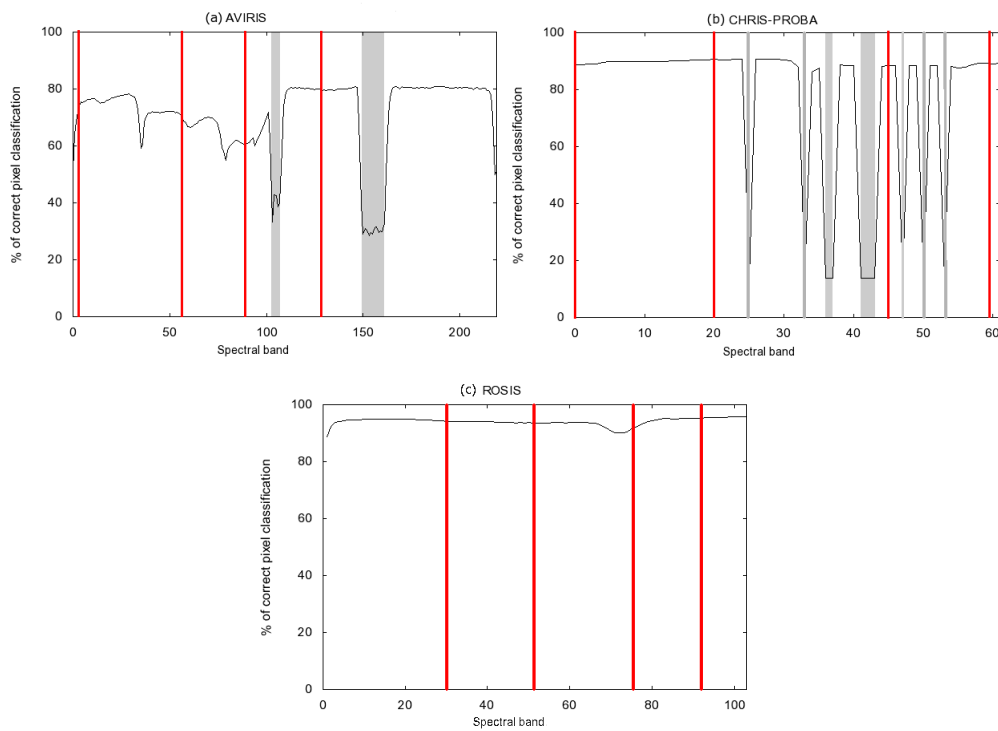
Besides, two different classifiers are used:

- K-nearest neighbour with  $k = 3$ .
- Support Vector Machine with third order polynomial kernel.

### 3.3.2 Band selection issues

The datasets are reduced by using a band selection method. However, before carrying out further experiments considering sets of bands obtained in that way, here an experiment is performed to check the validity of the information contained in all bands for three databases. In this experiments, Gabor filters are used to obtain features over each individual band. Therefore, each pixel is characterized by a feature vector of  $M \times N$  elements, being the number of orientations  $N = 4$ . The number of scales due to the spatial size of the images are  $M = 6$  for AVIRIS database and  $M = 7$  for both CHRIS-PROBA and ROSIS data sets. The classification process follows the classification scheme 1 and uses a  $k$ -NN algorithm with  $k = 3$ .

Figure 3.8 shows the classification results obtained for each one of the 220 bands of the AVIRIS dataset, the 62 bands of CHRIS-PROBA and the 103 bands of ROSIS (the noisy bands were not available in this case). Note that the best overall accuracy using the spectral features derived from a single band were around 80%, 90% and 94% respectively for each database. This shows that the spatial information is really useful for such classification tasks. This experiment also verifies that the previously detected noisy bands in each database (marked with grey shadows in the graphs) provide the worst results. In the case of the ROSIS dataset, the image provided only contained the 103 non-noisy bands and, therefore, no grey areas were marked. Besides, the bands selected using WaLuMI for  $B = 4$  are marked with red lines as an example to check that the selected bands are not the ones providing the best classification rates by itself, but they are always among the best ones. Note also that the band selection method is an unsupervised process that tries to maximize the information provided by the selected group of bands as a whole, not individually.



**Figure 3.8:** Classification rates for individual bands using Gabor filters. (a) AVIRIS (b) CHRIS-PROBA (c) ROSIS at the University of Pavia. The range of water absorption and the low SNR bands have been marked in grey. Selected bands using WALUMI with  $B = 4$  are marked with red lines.

### 3.3.3 Classification results

A result overview is shown in Figure 3.9 for two datasets (AVIRIS and CHRIS-PROBA) for each method of characterization. Both filter banks (dyadic and constant width) are used. Gabor stands for Gabor filter using individual planes, Gabor complex for the method that uses complex planes and introduces intra-channel information and last, Opponent features for the opponent features adapted to pixel characterization, explain in Section 3.2. Notice that another curve named spectral features appears in the comparison. This is the case of using the spectral values without any spatial information,  $\lambda_{x,y} = \{I^i(x,y)\}_{i=1}^B$ . The comparison with this baseline aims at showing what can be achieved using the same amount of data, if no spatial method is applied. For all the characterization methods, the classification scheme used is the scheme 1 with a K-nearest neighbour classifier with  $k = 3$ . All known classes are included in the training in the proportion they appear in the dataset. Hence the training per class is proportional to the size of the class and there are no classes over-trained.

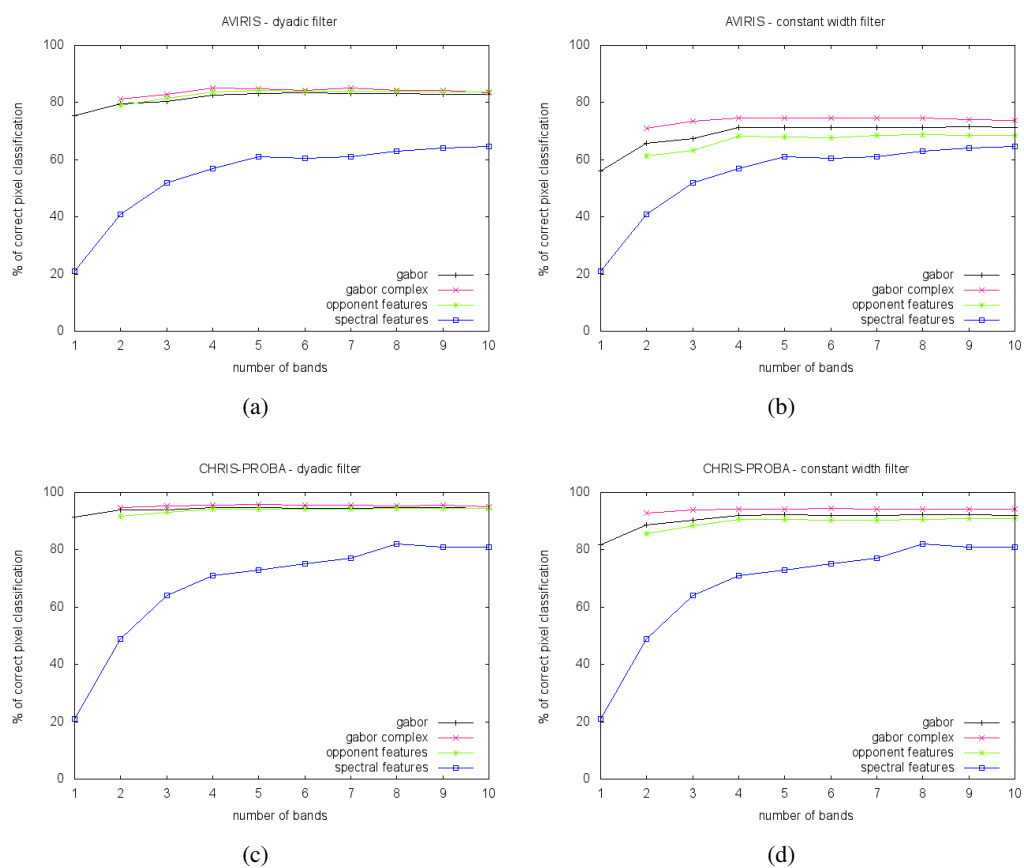
Figure 3.9 shows that for all methods, including spatial information allows to classify the image with a reliability that cannot be obtained if only the spectral information is used. This may be accomplished if more spectral features are used. However, it proves that spatial information allows to make more efficient use of data.

Regarding the different spatial-spectral methods, the differences are very small. Gabor with complex planes achieve a slightly better performance. Furthermore, the accuracy of the method with individual planes tends to stay above the one of the opponent features but notice that the differences are not significant. The difference between the type of filter used is relevant, a constant width filter bank performs worse than the traditionally dyadic filter bank. Differences are noticeable for AVIRIS and less significant for CHRIS-PROBA.

### 3.3.4 Scale analysis

In Figure 3.9 the pixel characterization vector is the response of each pixel to the entire filter bank, as described in Eq.( 3.2). The filter bank used is a Gabor filter bank so each filter in the bank has a certain spatial frequency scale and orientation. Consequently, within each characterization vector there is a number of features corresponding to the response to each of those filters. As mentioned before, each scale corresponds to different properties in the image, low spatial frequencies correspond to contributions of pixels to larger areas and higher frequencies are very characteristic of noise. When characterizing with all the features computed with the whole filter bank, this differentiation is not encountered. The target of this experiment is to analyze the scales independently to discern which frequencies are necessary and which are increasing the number of features needlessly, for characterizing landscape hyperspectral images.

In the first set of experiments, each learning curve is the result of classifying using the features computed with only the filters with the same scale. That is, 4 filters per scale in the case of Gabor over individual planes and opponent features and 8 for the complex plane method since the filters are not symmetrical in this case. In the case of the dyadic filters, the range of frequencies covered



**Figure 3.9:** Pixel classification rates for different characterization methods over AVIRIS and CHRIS-PROBA databases. (a)(c) Dyadic tessellation. (b)(d) Constant tessellation.

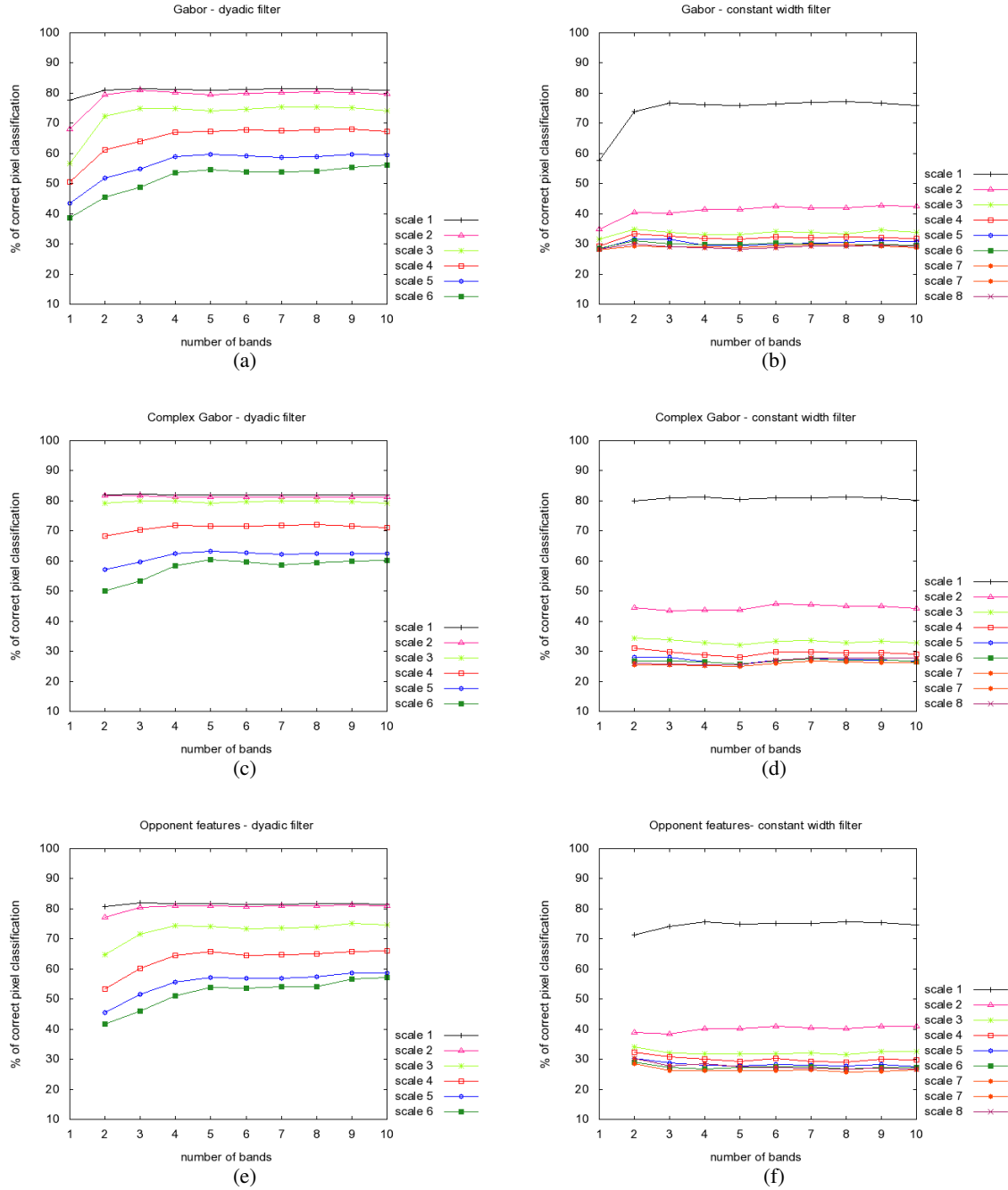
by each filter grow from lower to higher frequencies whereas for the constant width filters all filters will take an equal range of frequencies.

In Figure 3.10 and 3.11 each row shows the performance curves per scale for each characterization method. The left column shows the results when using a dyadic filter and the right one for a constant width filter. It is very relevant for this study that the curve corresponding to the lowest (first) scale always achieves the best performance, no matter which characterization method or filter type is used. Notice the bigger difference between filter types, this is due to the different size of the scale. One constant scale is equivalent to several dyadic ones at the lowest part since there the dyadic filter starts with smaller sizes and doubles the size later. For example, AVIRIS has a spatial size of  $145 \times 145$ , using 8 constant scales, where each of them includes a range of 18 frequencies whereas the four first scales of a dyadic filter have a size of 1, 2, 4 and 8. Hence, the first constant scale is approximately equivalent to the four first dyadic ones. Therefore, the big difference between the first and the rest when using a constant width filter is pointing out the relevance of the lowest frequencies in opposition to the rest which independent performance stays much lower.

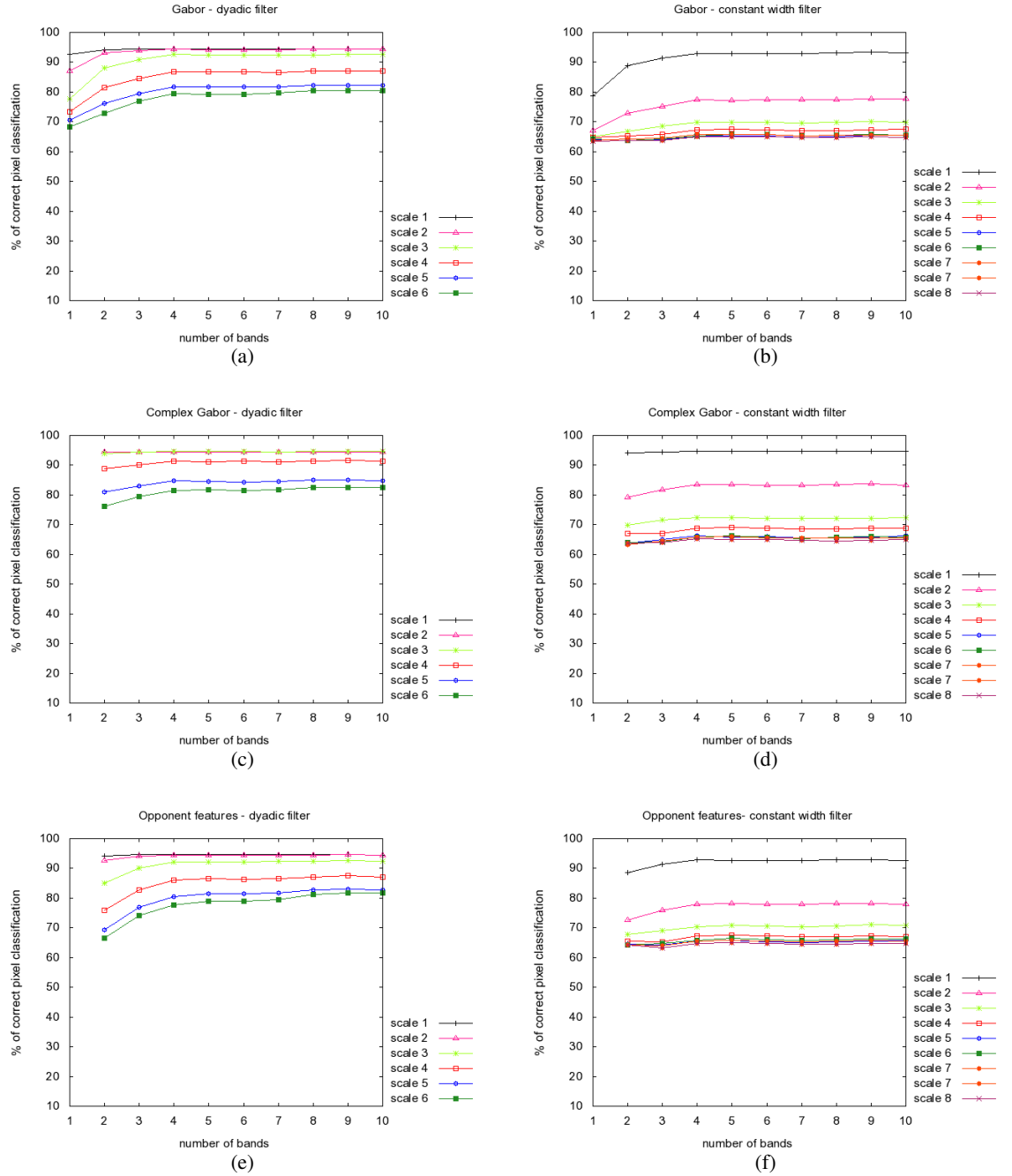
Results in Figure 3.12 and 3.13 are obtained progressively joining features. The first scale is used individually and the features from the next scale are joined one in each step progressively until the last one where all of them are already included. The fact that all performances curves stay together seems not to be informative but it is the opposite. Notice that the green curve (features from all scales) is always under or equal than the red one (joined features from scale 1 to scale 4), this means that including features from scales 5 and 6 did not lead to an improvement. Recall that by themselves, in the previous experiment, these scales were not providing significant results. This is confirmed by the results of the constant width filter (plots on the right). The first scale (the lowest) keeps staying above the rest. Notice that now the features are joined consequently the difference is not as big as when they were used independently. Although the features from the first scale are always included in the following but those still perform worse means that including more features than the first ones does not improve but also worsens the performance.

The last scale analysis consists of a spatial frequency descendant joining. Before the lowest scale was the starting point and higher spatial frequency scales were progressively joined. Figure 3.14 and 3.15 show the opposite joining, the highest scale is used independently and progressively next lower scales in opposite order are joined. This experiment is important to prove the importance of the lowest scales. The performance does not achieve the level seen before until lower frequencies are taken into account. As seen in the ascendant joining, including irrelevant features may worsen the performance and in this experiment the features from highest spatial frequency scales are always included.

These experiments show the significance of lowest scales in characterization and classification performance. Besides, it is noticed that including more features can have a negative effect. Remember that the datasets considered are all landscapes where big areas are observed, in terms of frequencies, that means that most of the relevant information is located in low and medium frequencies. Medium frequencies become important with the increase of the spatial resolution and the appearance of texture in the images. Figure 3.16 represents the results per method and filter type for sets of bands varying  $B \in [1 - 10]$ . The difference with Figure 3.9 is that now, according to the scale analysis performed, not all features are used. For the dyadic filter, features from scales

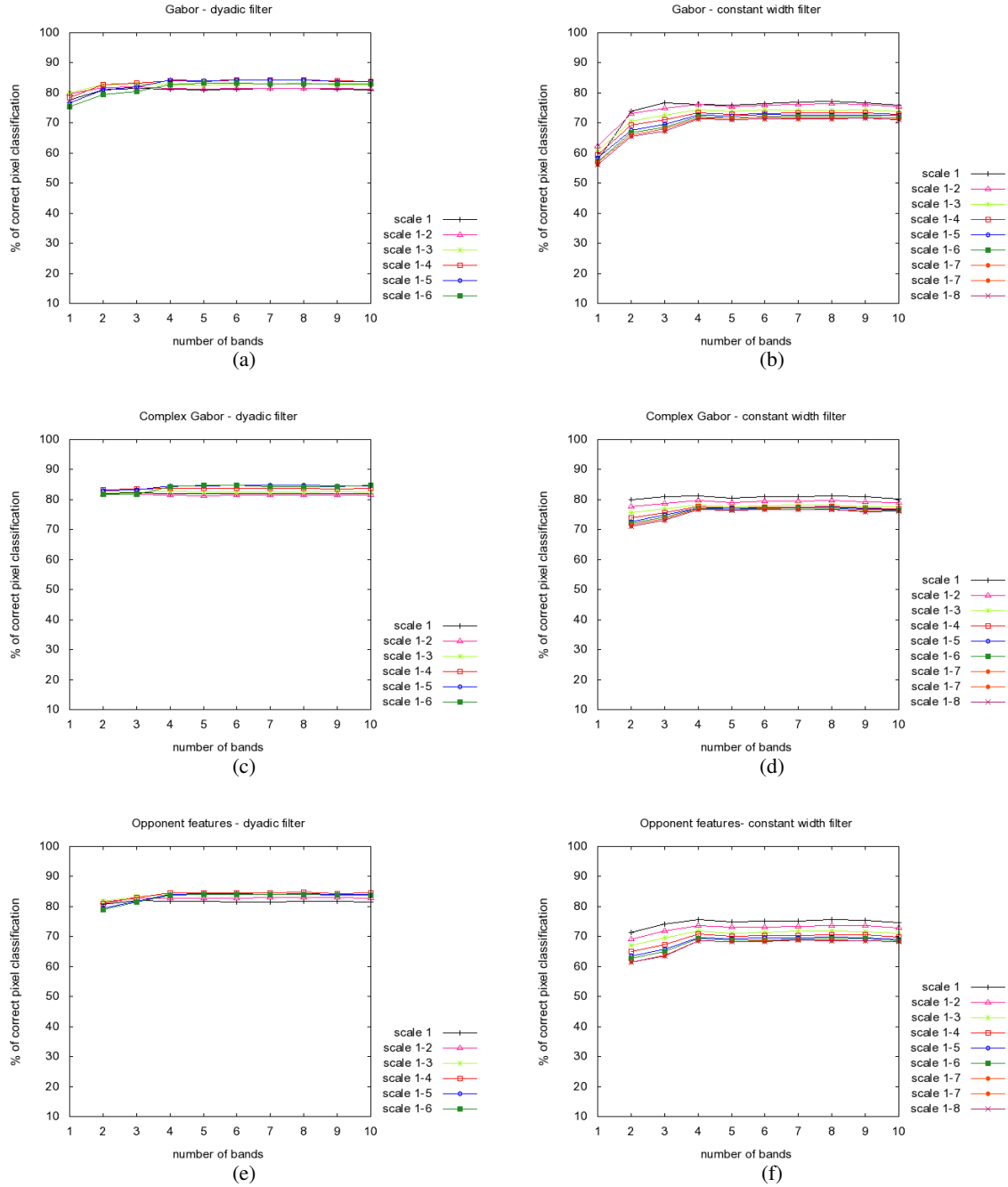


**Figure 3.10:** For the AVIRIS dataset, pixel classification rates using independently features from the same scale of the filter back. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.

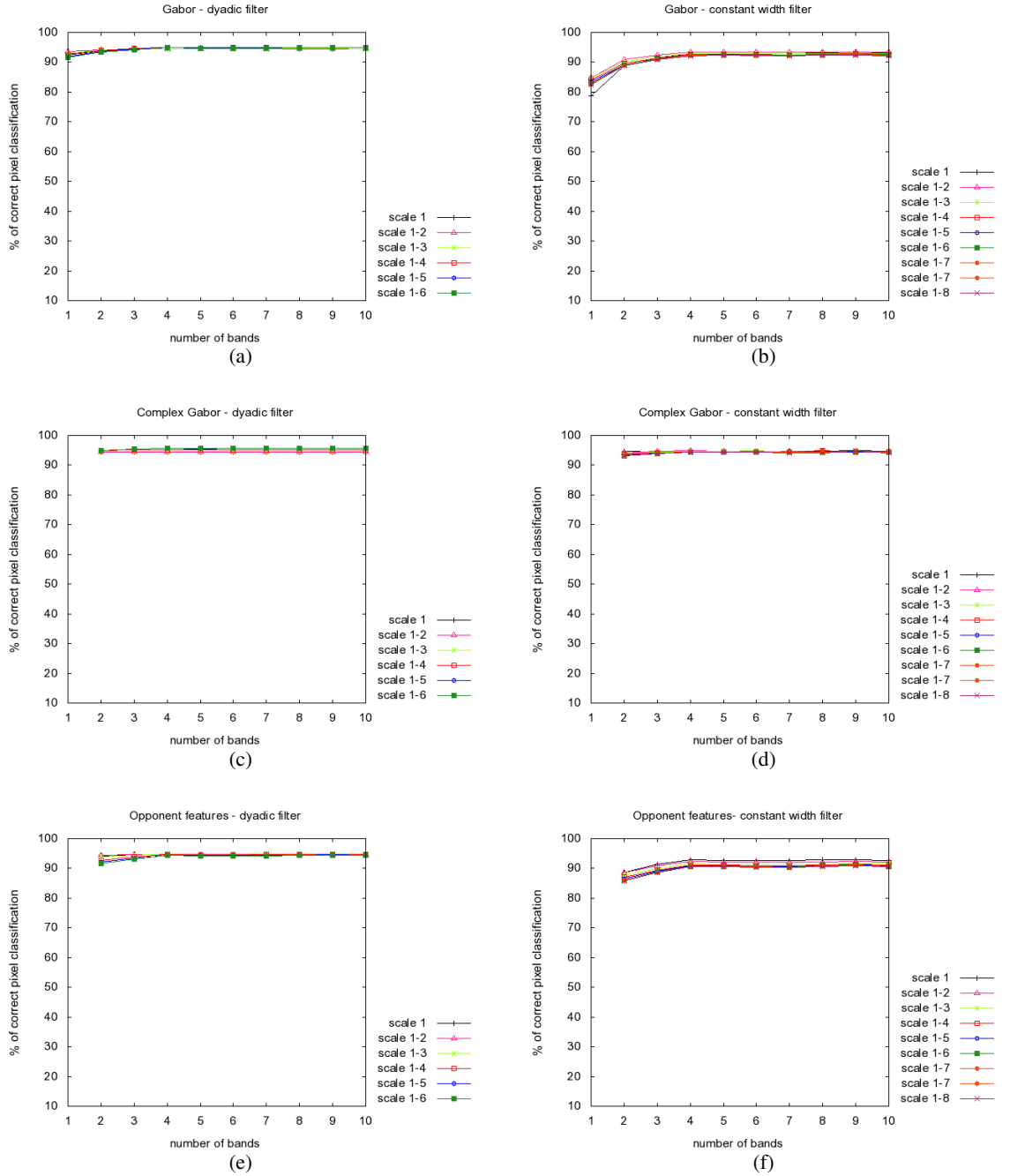


**Figure 3.11:** For the CHRIS-PROBA dataset, pixel classification rates using independently features from the same scale of the filter back. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.

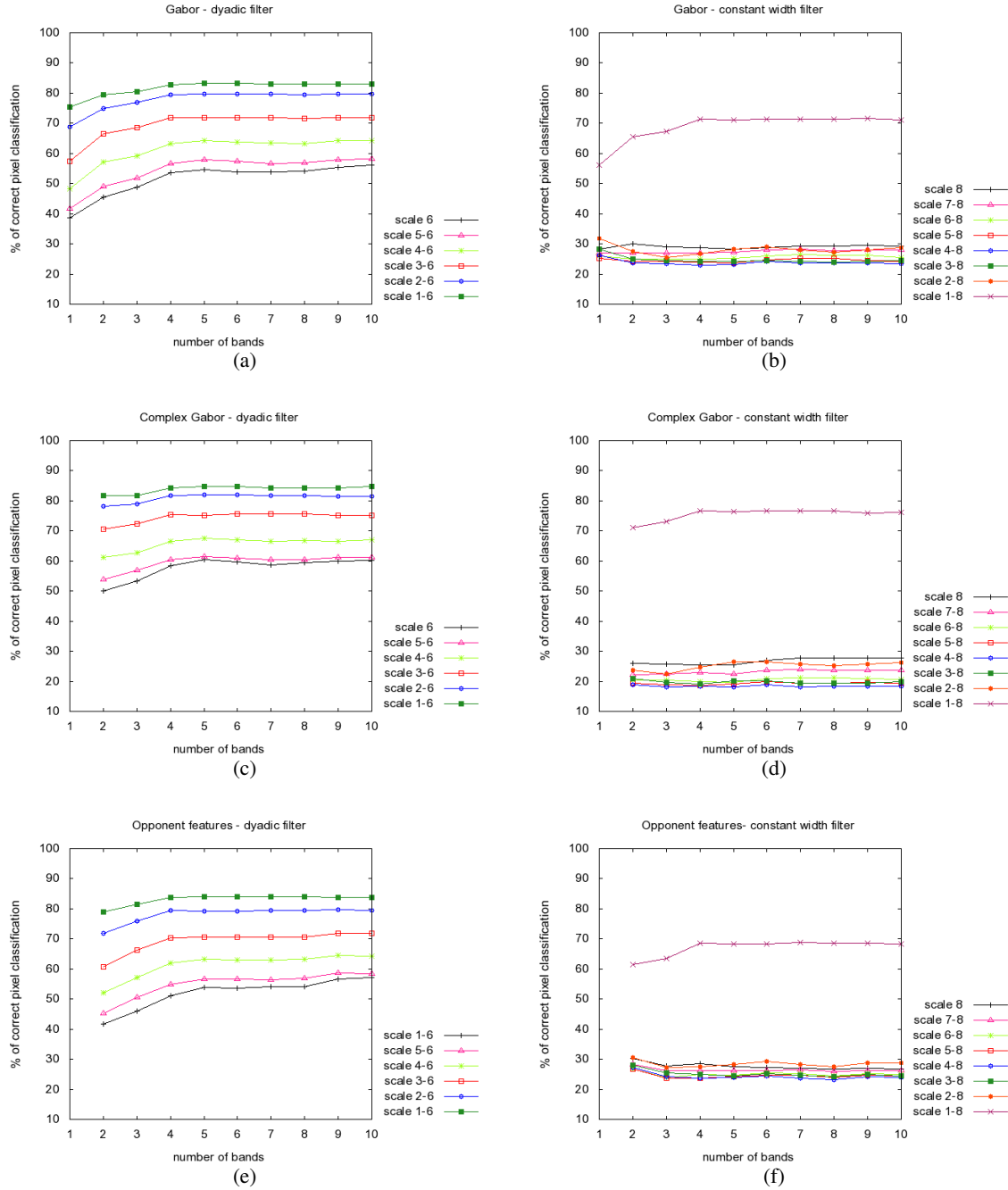




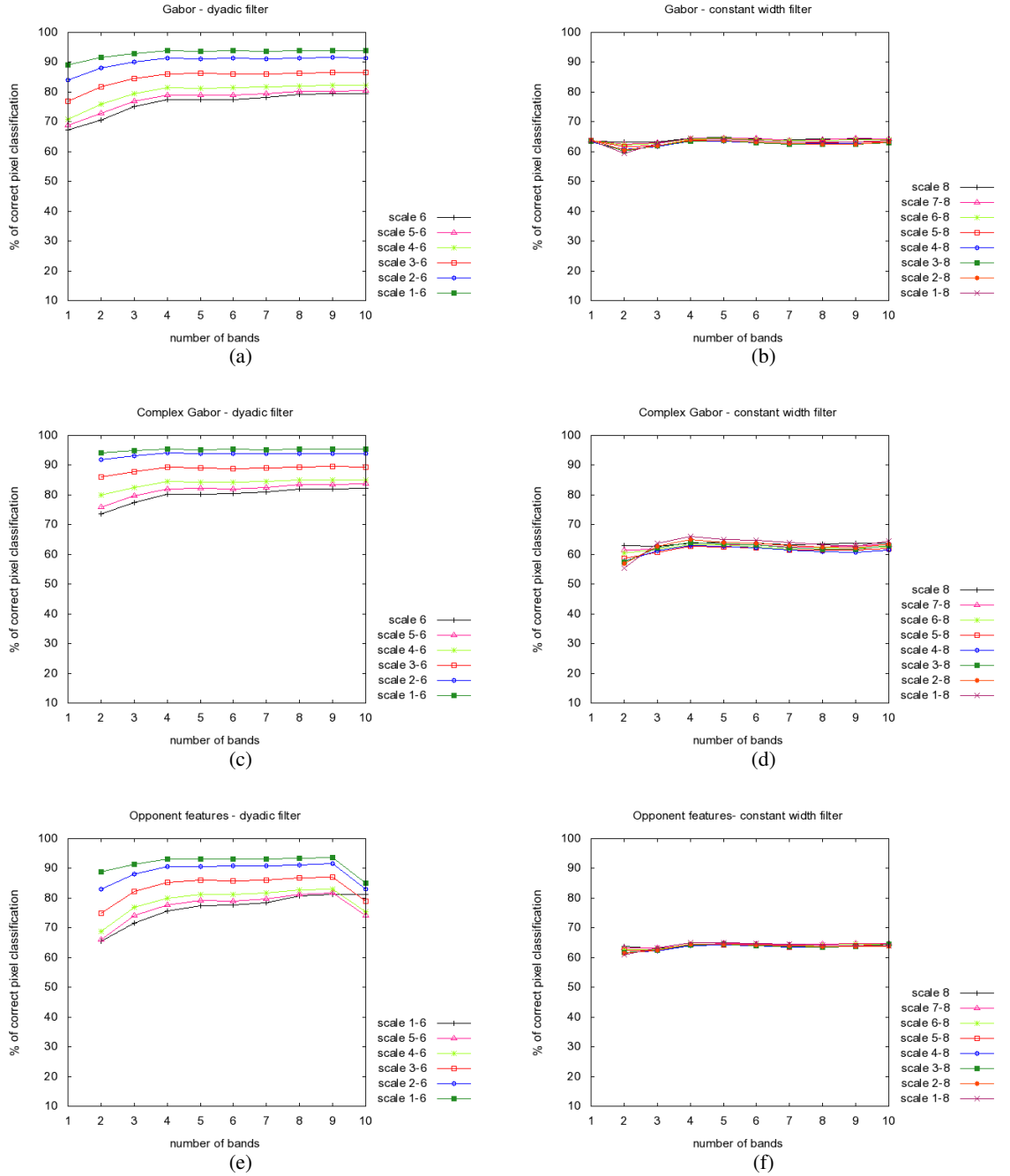
**Figure 3.12:** For the AVIRIS dataset, pixel classification rates using features starting from the first scale independently, and progressively joining the following scales. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.



**Figure 3.13:** For the CHRIS-PROBA dataset, pixel classification rates using features starting from the first scale independently, and progressively joining the following scales. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.



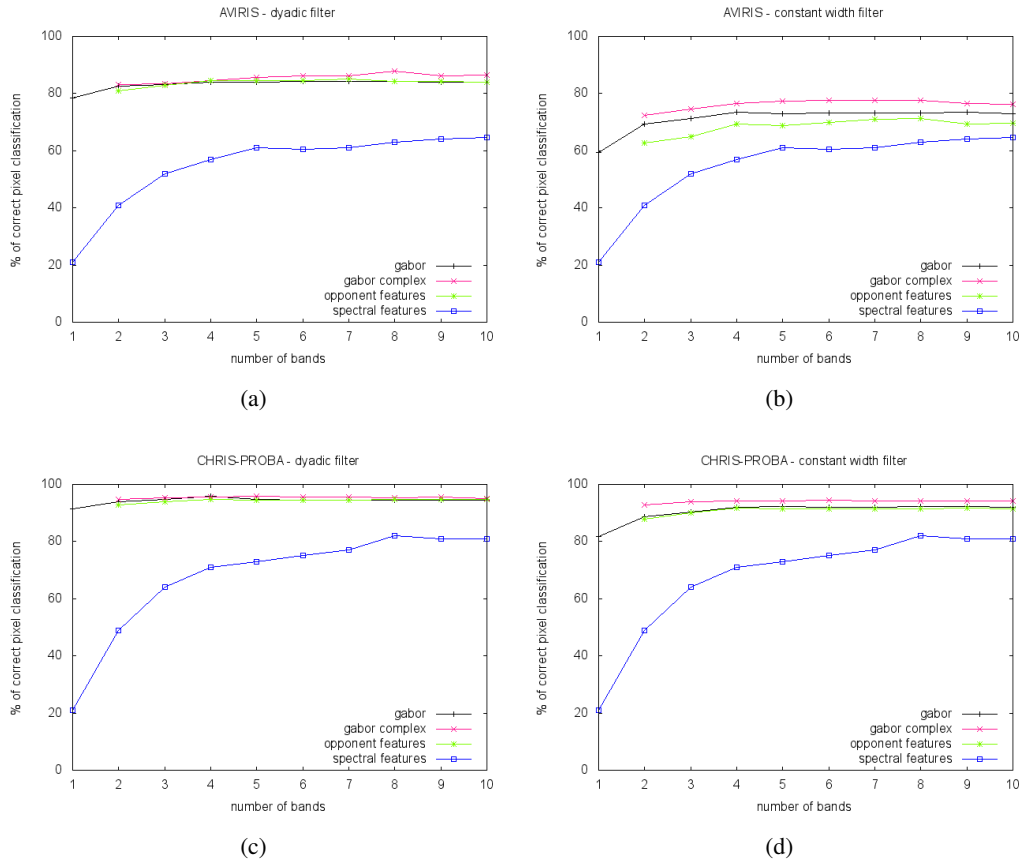
**Figure 3.14:** For the AVIRIS dataset, pixel classification rates using features starting from the last scale independently, and progressively joining the following scales from highest to lowest. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.



**Figure 3.15:** For the CHRIS-PROBA dataset, pixel classification rates using features starting from the last scale independently, and progressively joining the following scales from highest to lowest. Each different characterization method per row, Gabor over individual planes, Gabor over complex planes, opponent features, respectively. (Left) Dyadic tessellation. (Right) Constant tessellation.

$M \in [1 - 4]$  are included in the pixel characterization vector. As for the constant width filter set, features provided by the filters with  $M \in [1 - 6]$  are considered, like this, the spatial frequency range covered is equivalent to the dyadic one.

The differences between methods are reduced significantly and the performance is increased for AVIRIS dataset and maintained for CHRIS-PROBA. However, notice that in both cases the number of features is significantly reduced. Differences between dyadic and constant width filters are insignificant for CHRIS-PROBA but not for AVIRIS. Although in terms of range of frequencies both filters analyze the same range of frequencies, the dyadic filter analyzed the lower in detail whereas the constant does it equally. This may not be very important in CHRIS-PROBA because the classes presented are balanced. On the contrary AVIRIS has highly unbalanced classes and the detailed analysis helps to distinguish them.



**Figure 3.16:** Pixel classification rates for different characterization methods over AVIRIS and CHRIS-PROBA databases using a reduced number of features according to scale analysis performed. (a)(c) Dyadic tessellation. (b)(d) Constant tessellation.

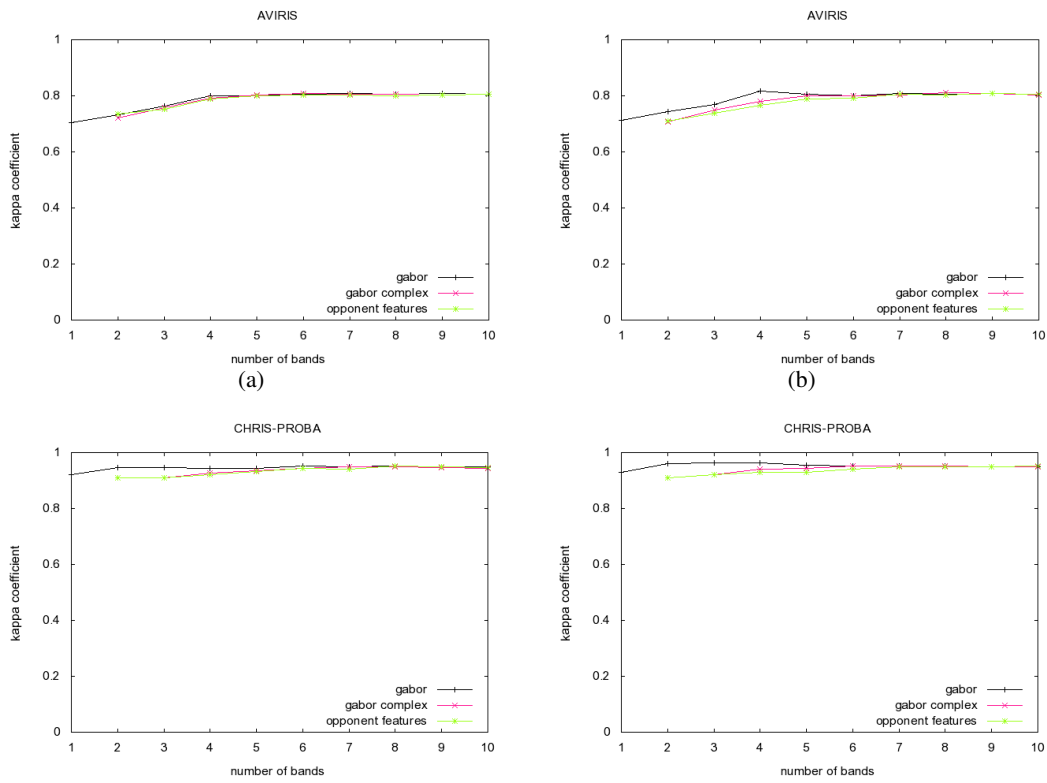
### 3.3.5 Segmentation

This section analyzes the results per class allowing to find out the behavior of the method regarding the small classes present in the datasets and considering all known and unknown classes to get a full segmentation of the image. The experiments are carry out using a SVM with a third degree polynomial kernel. SVM has been widely used in the field [43] [30] [107]. As in literature, the dataset is divided directly in training (5%) and test, which corresponds to classification scheme 2. Furthermore, only a dyadic bank filter is used. Constant width filters have been useful to analyze the importance of the different spatial frequency ranges. However, dyadic scales showed to perform better in the unbalanced case and on the other hand presented no disadvantages. Thus, standard Gabor filters with dyadic scales are chosen to show the segmentation results.

Until now, results were shown in overall accuracy and for the known classes. This was convenient for providing an overview and making conclusions in reference to overall accuracy and the role of the different scales. In this part, segmentation results are discussed, that allows to visualize the results and study per class accuracy which is interesting in the case of unbalanced datasets. When a dataset is highly unbalanced the overall accuracy may be biased by good results in bigger classes dismissing small ones. When segmentation is considered, dismissing small classes is not recommended because bigger classes will overtake the small ones making them disappear from the representation and the result will not represent the real landscape. Besides, considering only known classes does not obtain a whole segmentation of the image.

Because we are interested in the results per class, a first overview of the per class result can be given by substituting the overall accuracy of Figure 3.9 for the kappa coefficient [38][40]. Recall that the training selection is randomly made and a priori probabilities of classes are respected. This guarantees that all classes are trained proportionally and the kappa coefficient and per class accuracy given later are representative of the performance. Literature usually considers the classification of the known classes, as done here in all the previous experiments. However, a segmentation should show the whole image. For that purpose, in the segmentation results we consider a class called Heterogenous background that includes all the samples with unknown class. As the training also contain samples of this class randomly picked in the same percentage as the rest, the heterogeneity in it is also represented in the training set. The classification in this case can be considered as a classification problem where we try to find which samples belong to certain classes and which do not. This background represents a class which does not belong to the targeted classes.

In Figure 3.17 the kappa coefficient is plotted against the number of bands when all scales of the filter bank are used in the characterization. Remember that, in this case, the best performance for AVIRIS was 83%. This result was stable for  $B > 3$ . Notice that those results were given in overall accuracy and they may had been biased by the results of bigger classes. The kappa coefficient has not such a disadvantage and an increase on it is a signal of an improvement in all classes which makes the result faithful to small classes too. The curves in plot of Figure 3.17 show an improvement of about 10% between the first result where  $B = 1$  and the point where the result stabilized  $B = 4$ . This means that the classification becomes better in terms of all classes which is meaningful for small classes since their classification is a challenge. The difference is not that large when using CHRIS-PROBA dataset because their classes are not highly unbalanced.



**Figure 3.17:** Kappa coefficient against number of bands used for the different characterization methods over AVIRIS and CHRIS-PROBA databases using a number of bands  $B \in [1 \dots 10]$  and (a) a complete dyadic filter and (b) only with  $M = 1, 2, 3, 4$ .

The following results are organized to visualize the previous figures in terms of segmentation. Figure 3.17 shows the evolution of the results when the number of bands increases and features are extracted from a complete filter bank. The results stabilized for sets of bands bigger than four,  $B > 4$ . Hence, the segmentation results are presented using  $B \in [1...4]$  because it is the range where the performance and kappa coefficient increase and differences can be expected to appear. In Figure 3.17 no differences were found between the characterization methods. Besides, the number of features for Gabor using individual planes is lower than the rest, see Figure 3.6. Thus, while the performance in all perspectives is not worsen, the size of the feature vector is the lowest and prevents the danger of the curse of dimensionality. Consequently, segmentation results are presented using only the Gabor filter with individual planes characterization method.

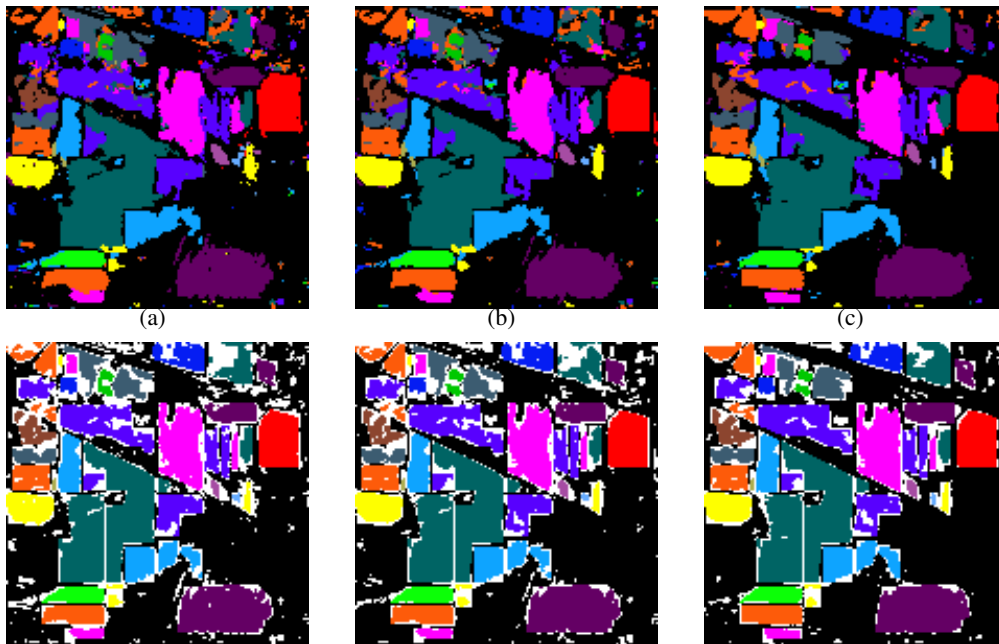
In Figure 3.18 the segmentations result of AVIRIS dataset is shown. Whereas the first row is the direct segmentation result, the second row represents all misclassified pixels in white so the error can also be easily detected. Notice that the error is localized in the spatial borders of the classes causing a more defined segmentation when the error decreases. No salt and pepper segmentation noise is observed, this is due to the usage of the filters that smoothes the characterizing of the pixels by considering frequential ranges, this smoothing characterization gives as a result a classification that tends to group neighborhoods of pixels avoiding possible noise in the characterization.

In Table 3.2 per class accuracy is presented per each experiment corresponding to the images in Figure 3.18. Regarding small classes like Stone-steel towers (95 pixels size), Alfalfa (54), Grass/pasture (26) and oats (20), only the smallest class (Oats) reports a result close to 0.5 which is still larger than the random classification. The rest of these small classes have a considerably good result taking into account that the training set contains respectively, only 4, 2, 1 and 1 samples of these classes. Compared with other methods [107][108][65], the results achieved here are comparable although those methods do not guarantee the proportionality of classes in the training set over training the small classes and use all the spectral information contained in the image.


















The scale analysis points out the possibility that, for this type of images, higher frequencies may include useless information within the characterization due to the nature of the images. This is accentuated in bigger classes and less important for the smaller ones. Notice in Figure 3.17 that the kappa coefficient stays in the same ranges for both cases, using the whole filter bank or reducing it to  $M = 1, 2, 3, 4$ , although the overall accuracy increases as seen in Figure 3.9 and Figure 3.16. As reported by Table 3.3 in comparison with Table 3.2, this is due to an increase in the performance of bigger classes whereas the small classes accuracy remains approximately the same. This is an advantage because by decreasing the information used the performance of big classes can be improved whereas the challenging small ones can remain in a considerably good performance with a proportional small training set.

Despite Figure 3.9, Figure 3.16, Table 3.3 and Table 3.2 reported results with lower and higher number of frequential scales, the progress of the increase could not be appreciated. Figures 3.20, 3.21 and 3.22 show the specific gain obtained by decreasing the number of frequential scales used in AVIRIS, CHRIS-PROBA and ROSIS datasets respectively.

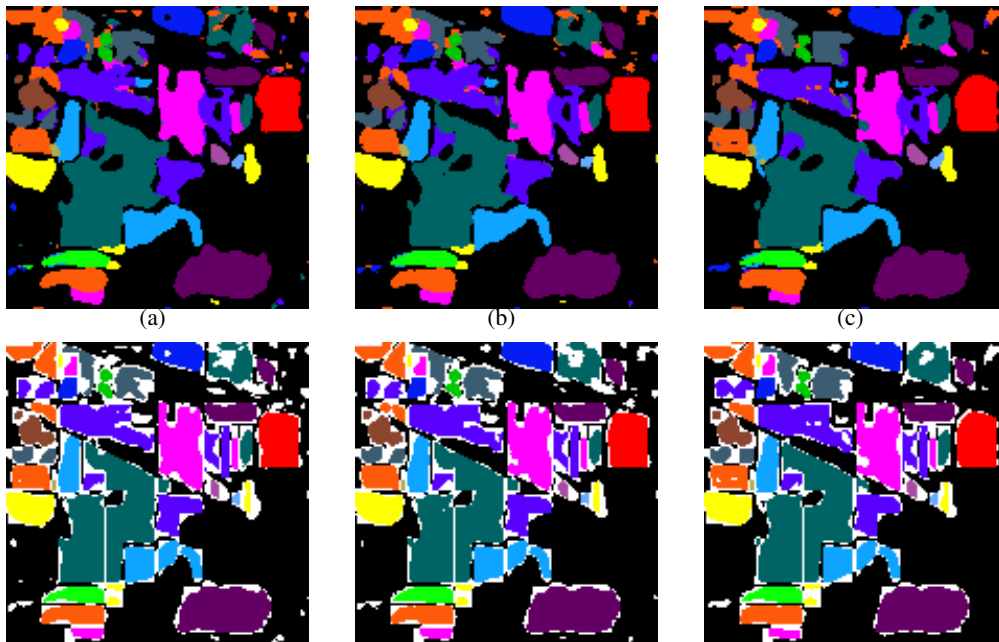





















**Figure 3.18:** Segmentation results for AVIRIS dataset using a complete filter bank and a different number of bands  $B \in [2..4]$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above.

classes		number of bands		
		2	3	4
Heterogenous background		0.8332	0.8642	0.8993
Stone-steel towers		0.6333	0.6111	0.6000
Hay-windrowed		0.9828	0.9699	0.9634
Corn-min till		0.7664	0.7841	0.8245
Soybeans-no till		0.8292	0.8270	0.8411
Alfalfa		0.4314	0.5686	0.6078
Soybeans-clean till		0.7118	0.7221	0.7530
Grass/pasture		0.7209	0.7125	0.7040
Woods		0.8772	0.8886	0.9089
Bldg-Grass-Tree-Drives		0.7091	0.7202	0.7396
Grass/pasture-mowed		0.4800	0.4400	0.4400
Corn		0.6441	0.6261	0.6892
Oats		0.5263	0.5263	0.5263
Corn-no till		0.7858	0.7997	0.8136
Soybeans-min till		0.8469	0.8512	0.8951
Grass/trees		0.8265	0.8378	0.8646
Wheat		0.8905	0.9204	0.8856
AA		0.8220	0.8412	0.8708
kappa		0.7536	0.7784	0.8181
AA without background		0.8106	0.8176	0.8416
kappa without background		0.9256	0.9252	0.9463

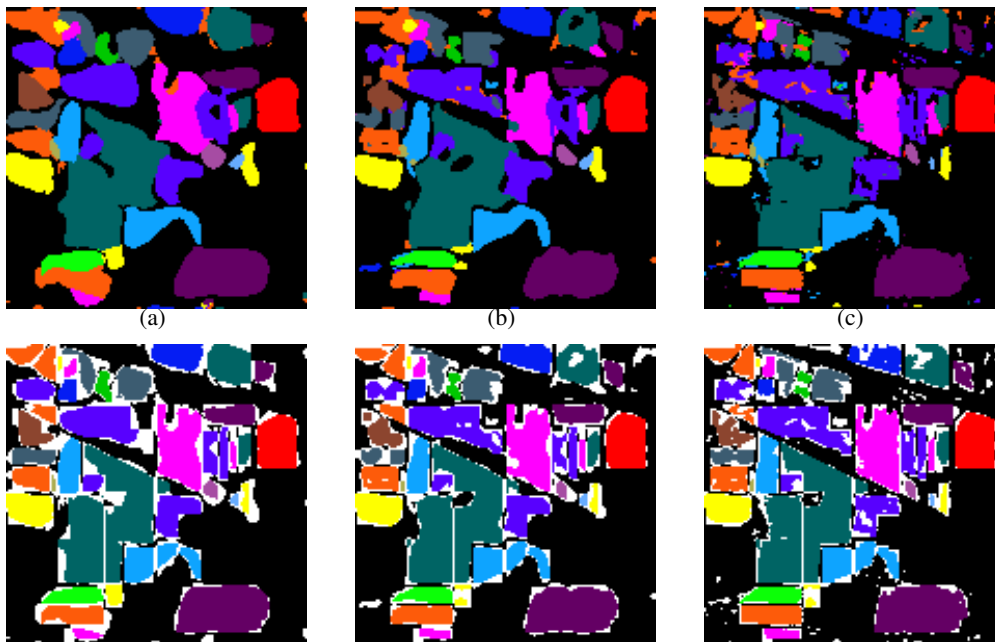
**Table 3.2:** Accuracy per class for the 17 classes classification of the AVIRIS dataset using the complete filter bank and different number of bands.




















**Figure 3.19:** Segmentation results for AVIRIS dataset using a filter bank with  $M = 1, 2, 3, 4$  and a different number of bands  $B \in [2..4]$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above.

classes		number of bands		
		2	3	4
Heterogenous background		0.8584	0.8728	0.8862
Stone-steel towers		0.6888	0.6888	0.6555
Hay-windrowed		0.9548	0.9419	0.9548
Corn-min till		0.8472	0.8270	0.8421
Soybeans-no till		0.8226	0.8150	0.7965
Alfalfa		0.5098	0.5490	0.6274
Soybeans-clean till		0.6415	0.7118	0.7478
Grass/pasture		0.7589	0.7589	0.7484
Woods		0.8479	0.8585	0.8691
Bldg-Grass-Tree-Drives		0.8947	0.9141	0.9168
Grass/pasture-mowed		0.7200	0.7200	0.7600
Corn		0.7297	0.7027	0.7117
Oats		0.4736	0.4736	0.4736
Corn-no till		0.7872	0.7872	0.8217
Soybeans-min till		0.8673	0.8618	0.8912
Grass/trees		0.8363	0.8293	0.8265
Wheat		0.8208	0.8606	0.8159
AA		0.8413	0.8495	0.8633
kappa		0.77779	0.7886	0.8077
AA without background		0.8237	0.8255	0.8397
kappa without background		0.9541	0.9523	0.9586

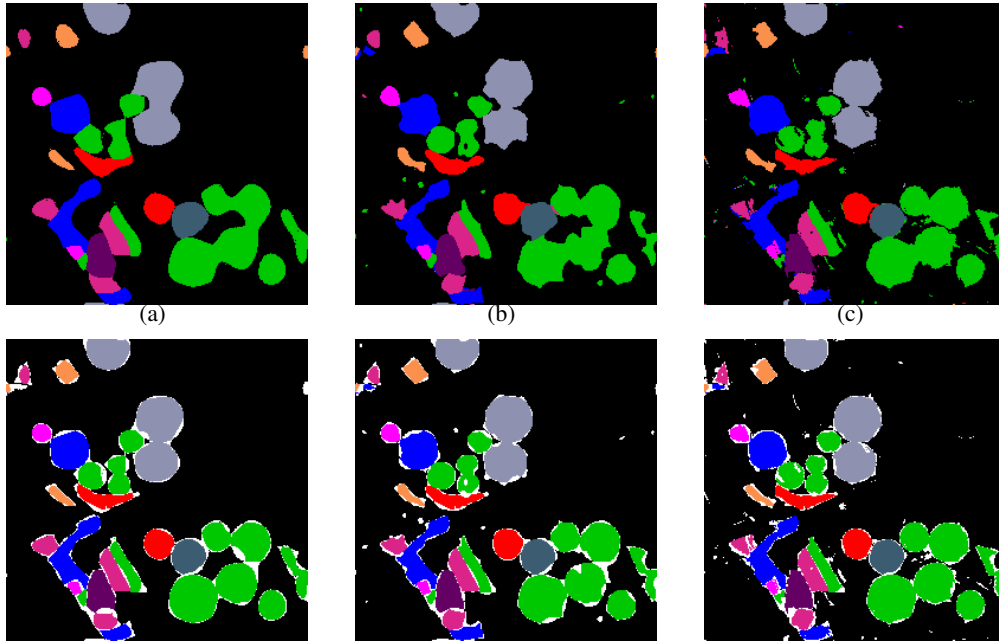
**Table 3.3:** Accuracy per class for the 17 classes classification of the AVIRIS dataset using a filter bank with  $M = 1, 2, 3, 4$  and different number of bands  $B \in [2..4]$ .













**Figure 3.20:** Segmentation results for AVIRIS dataset using  $B = 4$  and a filter bank with (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above.

classes		scales		
		2	4	6
Heterogenous background		0.8379	0.8862	0.8993
Stone-steel towers		0.8222	0.6556	0.6000
Hay-windrowed		0.9290	0.9548	0.9634
Corn-min till		0.8422	0.8422	0.8245
Soybeans-no till		0.7791	0.7965	0.8411
Alfalfa		0.8824	0.6275	0.6078
Soybeans-clean till		0.7925	0.7479	0.7530
Grass/pasture		0.8541	0.7484	0.7040
Woods		0.9033	0.8691	0.9089
Bldg-Grass-Tree-Drives		0.9058	0.9169	0.7396
Grass/pasture-mowed		0.7200	0.7600	0.4400
Corn		0.7658	0.7117	0.6892
Oats		0.6842	0.4737	0.5263
Corn-no till		0.8511	0.8217	0.8136
Soybeans-min till		0.8913	0.8913	0.8951
Grass/trees		0.8350	0.8265	0.8646
Wheat		0.8955	0.8159	0.8856
Acc		0.848360	0.863329	0.8708
kappa		0.791008	0.807732	0.8181
Acc without background		0.859159	0.839764	0.8416
kappa without background		0.965104	0.958606	0.9463











**Table 3.4:** Accuracy and kappa per class for the 17 classes classification of the AVIRIS dataset using  $B = 4$  and a filter bank with (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6$ .



**Figure 3.21:** Segmentation results for CHRIS-PROBA dataset using a filter bank with  $B = 4$  and (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6, 7$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above.

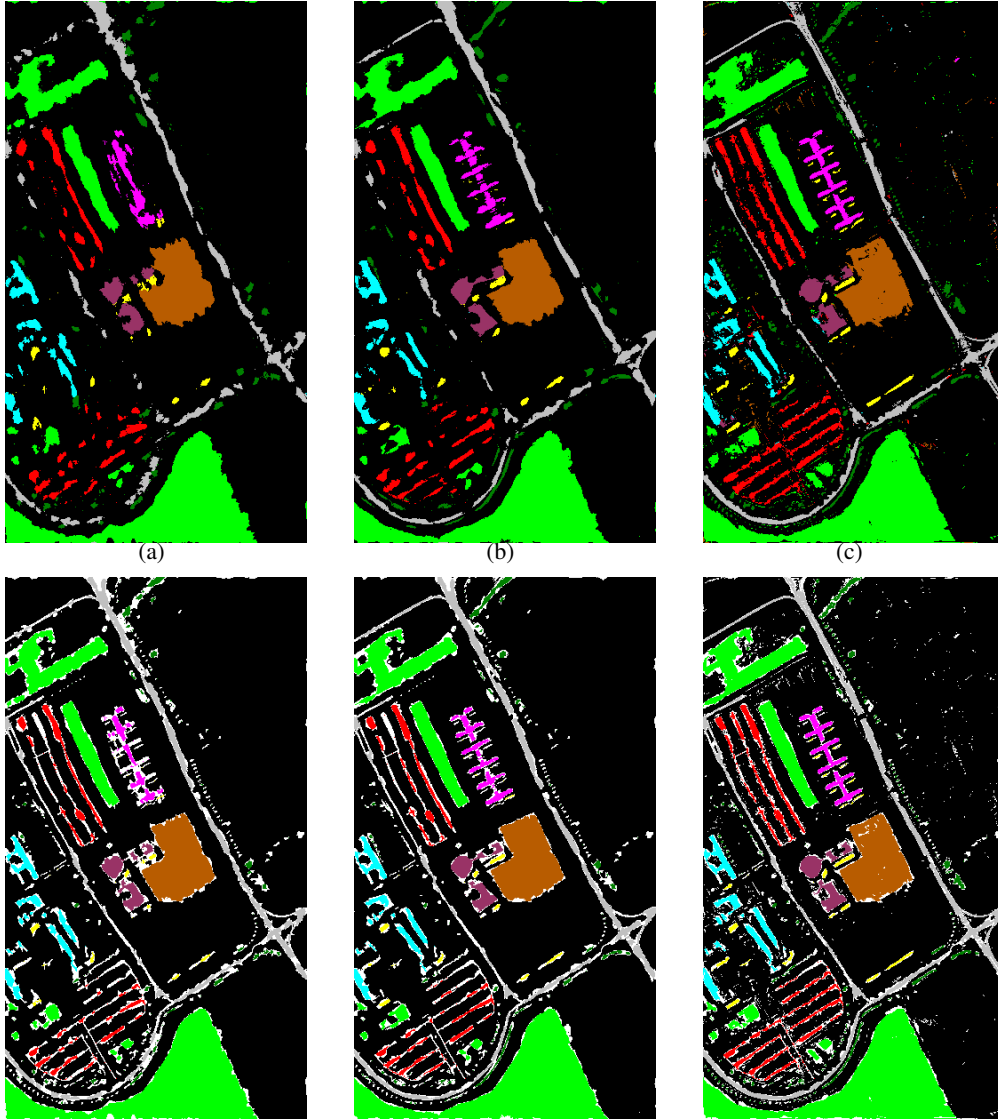
classes		number of frequential scales		
		2	4	7
Heterogenous background		0.9772	0.9740	0.9765
Pot		0.9008	0.8927	0.8757
Alfalfa		0.8887	0.8994	0.8795
Corn		0.8992	0.8888	0.8986
Garlic in greenhouse		0.8544	0.8606	0.7616
Grass		0.9085	0.	0.8612
Onion		0.8367	0.8484	0.8139
Garlic		0.9226	0.9211	0.9136
Sugar cain		0.8097	0.8340	0.7753
Sunflowers		0.9405	0.9496	0.9375
Acc		0.9570	0.9546	0.9538
kappa		0.8997	0.8942	0.8916
Acc without background		0.8986	0.8982	0.8879
kappa without background		0.9765	0.9801	0.9795

**Table 3.5:** Accuracy and kappa per class for the 10 classes classification of the CHRIS-PROBA dataset using  $B = 4$  and a filter bank with (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6$ .

classes		number of frequential scales		
		2	4	7
Heterogenous background		0.9551	0.9586	0.9407
Pot		0.6034	0.7071	0.7955
Alfalfa		0.9318	0.9410	0.9607
Corn		0.6644	0.6790	0.8094
Garlic in greenhouse		0.2288	0.3171	0.5206
Grass		0.5133	0.7973	0.9076
Onion		0.9579	0.9616	0.9472
Garlic		0.7508	0.7958	0.8520
Sugar cain		0.4318	0.5052	0.7449
Sunflowers		0.2614	0.3659	0.6640
Acc		0.9115	0.9239	0.9249
kappa		0.7447	0.7836	0.7995
Acc without background		0.7439	0.7904	0.8644
kappa without background		0.9374	0.9487	0.9573

**Table 3.6:** Accuracy and kappa per class for the 10 classes classification of the ROSIS dataset using  $B = 4$  and a filter bank with (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6$ .





**Figure 3.22:** Segmentation results for ROSIS dataset using a filter bank with  $B = 4$  and (a)  $M = 1, 2$ , (b)  $M = 1, 2, 3, 4$ , (c)  $M = 1, 2, 3, 4, 5, 6, 7$ . On the first row the segmentation result is shown. The representation of the error in white can be seen in the second row respectively for each segmentation result above.

### 3.4 Conclusions

A new schema for classification and segmentation of hyperspectral landscape imaging is presented. It pursues two goals: improving the classification results and decreasing the dimensionality. The traditional scheme uses the entire spectral signature of each pixel (possibly enriched with spatial information) to carry out per pixel classification and perform spatial corrections over the result. This increases the dimensionality of the problem and often results in over-segmentation.

We suggest a new scheme that starts using an unsupervised method for reducing the dataset, then spatial-spectral characterization is applied to replace the spectral vector traditionally used. Last, it performs per pixel classification providing the direct result, a classification segmentation map.

Three methods are presented for spatial feature extraction in hyperspectral pixel characterization. It is experimentally proven that the proposed scheme, with these characterization methods, provides good classification rates. This is remarkable in datasets with extreme unbalanced classes. Furthermore, the approach presented here uses a reduced set of selected spectral bands, simplifying the representation. This is important in order to avoid the problems caused by the curse of dimensionality and also because it leaves room for other features to be used to improve the characterization.

We show that the spatial information provides an appropriate characterization of the pixels for classification tasks. These features lead to good classification rates. We also show that the spatial information influences the characterization process much more than the inter-channel information. No big differences are found between the three sort of spatial features analyzed although they have big differences in the number of features used to describe each pixel, being the method proposed by applying Gabor filters over individual bands the most appropriate because of its simplicity and smaller dimensionality.

We also study the influence of the different scales in the feature extraction process and found that, when only smooth areas compose the image, the first scales provide the best characterization and the addition of the last scales tends to worsen the classification results. However, if we have to deal with non-homogeneous regions, the use of the medium scales may improve the characterization.

In the segmentation experiments, we find that most of the miss-classified pixels fall in the borders of the labeled regions where the spatial features can be confused due to the background information or due to the transitions between different classes in the image plane. However, the segmentation of the inner part of the regions is always remarkably homogeneous without needing any spatial post-regularization.

## Training selection

In this manuscript, we face segmentation and classification as a single problem by using pixel classification. Notice that for this task training data is needed. Training data is provided by expert labeling which in hyperspectral remote sensed images means a group of experts moving through a considerably large land extension. Consequently, labeling is expensive and reducing it is convenient. Some authors work in a supervised scenario where prior knowledge is available and training data is selected within each class, as we perform in previous chapter or in literature [107][30]. To do this, a previous knowledge of the whole dataset is need.

Active learning techniques have been also applied in previous studies [98][65]. In these, the expert collaboration improves progressively the training data and a complete knowledge is not needed a priori. This starts with a low amount of data and iterates in collaboration with the expert. Note that the interaction with the expert in different steps, in this specific case, means moving the experts and the equipment to take samples in land fields. This process is tedious and highly time consuming. In addition, in both cases (active learning and supervised classification), the training data is generally first selected by randomly selecting among all data. Although unsupervised, this may not be reliable. Randomly distributed samples can lie in non interesting areas and reducing the size of the starting training set may turn training data into non representative.

To face this problem the most interesting samples should be provided to the expert from the beginning. We aim at providing a solution to this concern by designing a technique to select more representative training data so the amount of data labeled can be reduced while maintaining the results shown up to now. This technique must be an unsupervised process to effectively decreased the expert collaboration.

In unsupervised scenarios, when no prior knowledge is available, data analysis techniques are widely used for finding relevant data. Among them, clustering techniques allow to divide data into groups of similar samples. A very large number of clustering techniques is available but some of them rely upon a prior knowledge, such as the number of clusters and the shape of clusters in the feature space (often elliptical). When dealing with an arbitrarily structured feature space, only non-parametric methods are suitable because no model assumption can be made [55]. Nonparametric

techniques are those that do not assume that the structure of the data is fixed. Typically, the model grows in size to accommodate the complexity of the data and assumptions about the types of connections among variables are made. In this sense, the methods can be distinguished into hierarchical and density based procedures. The first ones either aggregates or divides the data according to some agreed measure. Density based procedures consider the probability density function of the feature space and search for local maxima and based on the local structure of the feature space, a number of samples are associated to the maxima found [103].

Many clustering methods have been applied to image segmentation in various fields and applications [37]. However, fully unsupervised procedures often have insufficiently accurate segmentation results. For such a reason, a hybrid scenario between supervised and unsupervised techniques is of our interest. In this case, the methods applied count with some labels to train a classifier but a complete prior knowledge of the data is not needed. Because labeling is neither fast nor cheap, the fewer labeled data needed, the least the experts need to collaborate [74].

In this chapter a new unsupervised method is suggested for selecting the samples of the training set. It consists on using a clustering analysis to find samples of interest. This technique is included in the scheme suggested in the previous chapter. After selecting the most suitable data for training, these samples are labeled by the expert and used to train a classifier completing the classification-segmentation task. Notice that because the selection provides a set of the most interesting samples for training, this data is more representative than the one selected at random and thus less data is needed to achieve the state of the art results. Hence, although the data has to be labeled by experts, their collaboration is reduced. The experts provide the classes of the samples selected, they do not provide prior information. The number of classes is an unknown factor solved by the unsupervised selection method. For this reason we claim that this scheme (once added the selection scheme) is a semi-supervised technique.

## 4.1 Training selection

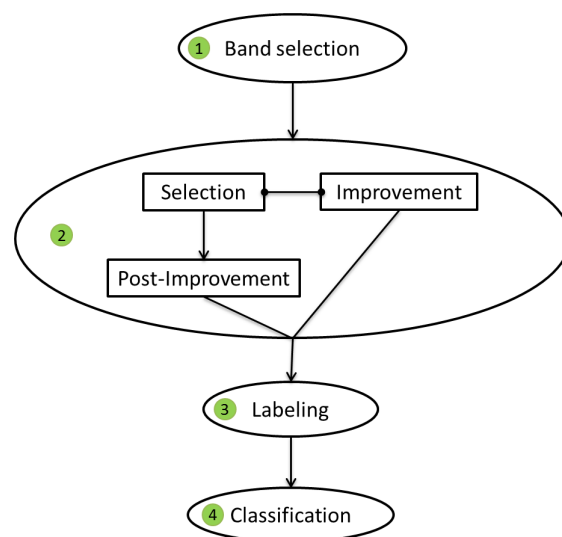
Comaniciu et al. states in [22] that vision tasks can be improved if they are supported by more reliable data. Nowadays databases used for segmentation and classification of hyperspectral satellite images are fairly reliable in terms of spectral and spatial resolution. Therefore, we can consider that the representation of the data is reliable. However, training sets are often built by randomly picking a percentage of samples.

Notice that hyperspectral land scape images contain different types of fields. Even if some of them belong to the same type of field they are located on a different type of soil or humidity conditions which can make them slightly different. In these cases it would be convenient to have samples from different conditions although they belong to the same class. When random selection is performed the significance of the samples picked for training can not be guarantee, they may lie all over the image and samples for different areas of the same class can be missed.

One of the main contributions of this thesis is the introduction of a unsupervised method to select the training data. The novelty of this suggestion is the idea that the training can be chosen, in an unsupervised way, to be representative enough of the data. The algorithm makes an unsupervised

selection of the training samples based on the analysis of the feature space using non-parametric clustering. Then, these samples of interest are labeled by an expert and used to train a classifier. This method guarantees that without any previous knowledge the training is free of redundancies and representative. It proceeds as follows:

1. A band selection method is used. With it the data set is reduced to a smaller set of bands. This set is less correlated than the original while it keeps as much information as possible. We used the WALUMI band selection method [72], but any other band selection method that fulfills that requirement can be used instead.
2. A non-parametric clustering technique is used and prior knowledge is not needed. The clustering procedure is applied over the reduced dataset. The centers of the clusters found form the selected training set.
3. The expert is involved once, after the selection, to provide the corresponding labels of the selected samples. We simulated the expert by checking the corresponding labels on the groundtruth available.
4. A classifier is built using the training set defined before. Although the clustering is performed using spectral features, we test that the selection obtained can be used independently to the type of features used for classifying.



**Figure 4.1:** Diagram representing the flow of the new classification scheme that includes the training selection. Each step is numbered as seen in the text. Notice that the selection can be performed with internal improvements or with post-process improvements.

Figure 4.1 shows the flow diagram of the scheme presented. Observe that the second step (selection) can introduce improvements in two ways, as a post process or within the selection itself. In the following sections the selection will be introduced and will be followed by the explanation of improvements specifying if these are within the selection algorithm itself or a post-process after the selection.

#### 4.1.1 Mode seeking clustering

An unsupervised technique is needed to perform an analysis of the data. We want our method to be unsupervised so that it does not use prior knowledge for the selection. Consequently, those clustering methods where the number of clusters has to be stated are discarded. Among the rest, mode seek [28] was chosen for being a rather general clustering approach.

Mode seeking clustering is a well known clustering principle for image segmentation. Based on a given set of objects, in case of images these are pixels, a non-parametric estimate of the probability density function (pdf) is made. The modes of this pdf correspond to the clusters. In a gradient search all objects are used as a starting point and objects ending up in the same mode belong to the same cluster. Neither the number of clusters nor their shape has to be predefined.

The most popular mode seeking procedure is the mean shift algorithm [18][41]. It is based on a Parzen kernel density estimate of the pdf. In contrast to the classic K-means clustering [27], or the more advanced Mixture-Of-Gaussian density estimates there are no embedded assumptions on an underlying Gaussian distribution of the data [18] [22]. In the mean shift algorithm the direction of the local gradient is found by a shift of the mean of the local mean when the distances to the objects in a local neighborhood are weighted by the chosen kernel. This procedure works well for the segmentation of color images, especially when some spatial information is included in features representing the pixels [22]. Problems with mean shift are that the modes as well as the convergence are not sharply defined. Thereby, separate nearby modes may be found that are erroneously not merged. Moreover, formally all pixels have to be used as a starting point, which is very time consuming.

Another algorithm based on mode seeking is  $s$ -NN mode seeking. Instead of the Parzen kernel density estimate, this method is entirely based on the distances to the  $s$ -th neighbor. It can be traced back to a proposal by Koontz et al. in 1977 [60]. It has been in the Matlab toolbox PRTools [28] for around 20 years. Recently it has been redefined [29] and compared with mean shift. The procedure can be summarized as:

Do for all objects:

1. Find its  $s$  nearest neighbors.
2. Use the distance to the  $s$ -th neighbor as a measure for the density (in fact one over the distance).
3. Define a pointer to the object with the highest density in the  $s$ -neighborhood.
4. From all objects follow the pointers reaching objects that point to themselves: the modes.

Various implementations are studied. We used one that is based on an approximate nearest neighbor search [5]. It performs the above algorithm for clustering 10366 objects in 5 dimensions with  $s = 100$  in 1.4 seconds and with  $s = 10$  in less than a second (0.7) on a standard PC (Intel Core Duo 2GHz, with 4GB of RAM). Its computational complexity is about  $O(sn^2)$  for data sets with  $n$  objects. The dependency on the dimensionality is heavily problem dependent due to the approximate nearest neighbor. Advantages of this algorithm over mean shift are that it is much faster and converges exactly to modes that correspond with objects (pixels). Moreover it can handle high dimensional spaces and finds solutions for sets of  $s$ -values in almost the same time as needed for the largest  $s$ -value in the set.

Notice that  $s$  influences the number of clusters found following the rule: the bigger the  $s$  the lower the number of clusters and viceversa. Hence, to change the size of the training data it is only necessary to tune the value of parameter  $s$ .

### Adding spatial information

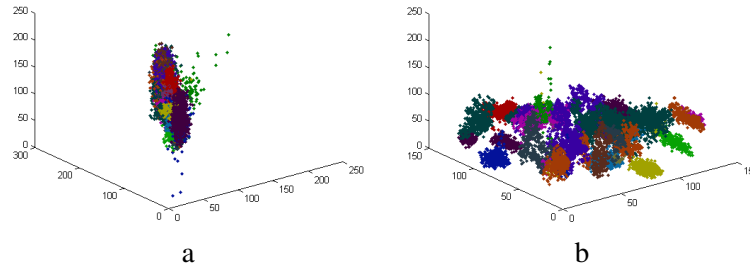
The nature of the feature space is application dependent. In the problem we tackle (land cover classification), our samples are pixels in the image space. Given an image there is a pair of spatial coordinates for each pixel in addition to all the spectral features given by the sensor. The improvement suggested here is based on two key points:

- Class connection principle or smoothness: spatially connected samples are likely to belong to the same class, that is, they are close in terms of spatial coordinates.
- Environment influence: when a class is located in more than one spatial location, even being the same class, the characteristics of their samples can differ due to different lighting or soil conditions in the different regions.

Thus, we suggest to incorporate spatial information to the selection algorithm for the sake of clustering pixels regarding also their spatial connectivity. This can be easily performed by adding the spatial coordinates to the feature vector of each pixel [78]. By adding a spatial component to the distance computation, samples nearby will have a higher probability of being clustered together and the opposite for spatially remote samples even if they belong to the same class.

The difference between samples is considered for the search of the local maxima. In that difference, all features (dimensions) are considered. When features do not include any spatial information the class connection principle is missed. By including the spatial coordinates among the feature of the samples we avoid to lose the class connectivity advantage. Remember that band selection is always performed before. Thus, the dimensions are those from the reduced dataset.

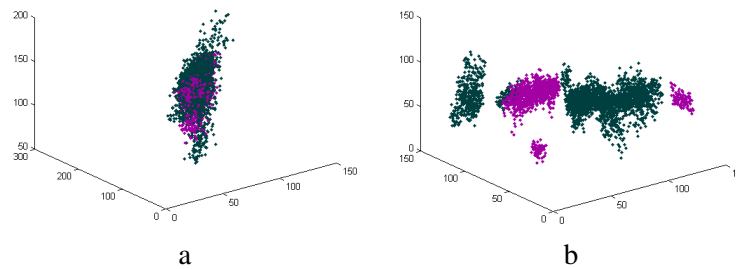
See Figure 4.2.(a) where all samples are represented in the three first features space and in a different color per class. In this representation no spatial data is considered. Note that all classes are located in the same space and, considering that no prior knowledge is available when clustering, finding representatives for each class is difficult. Moreover, different areas of the same class are within the same cloud. However, when spatial data is included, Figure 4.2.(b), the single clouds of samples split according to spatial distances and become more separable.



**Figure 4.2:** Three first features for all classes in the AVIRIS database (a) when no spatial coordinates are included (b) when spatial coordinates are included as features.

In Figure 4.3 the case of two classes of the image is studied, soybeans-no till (magenta) and soybeans-min till (dark green) (Figure A.1). Observe that they are composed of different areas located in different spatial locations. When no spatial information is considered in Figure 4.3(a) all of the samples lie in a similar area whereas when considering the spatial coordinates, Figure 4.3(b), these two clouds split in different groups corresponding to the different location in the space where the classes are found.

Notice that according to the groundtruth, Soybeans-min till has five locations and three of them are relatively near, this can be seen in Figure 4.3(b) as two clear groups of samples, one bigger and long. In the case of the Soybeans-no till, it is distributed along three places in the image and three different clouds are now visible. As the cluster detects those groups, a mode can be found and a training sample is selected for each area. One can already see an upcoming problem with neighbouring classes, unless they are big enough to be detected as their own cluster, they can be confused as part of the other one. This will be further discussed later.



**Figure 4.3:** Three first features for two classes in the AVIRIS database (a) when no coordinates are included (b) when spatial coordinates are included as features.

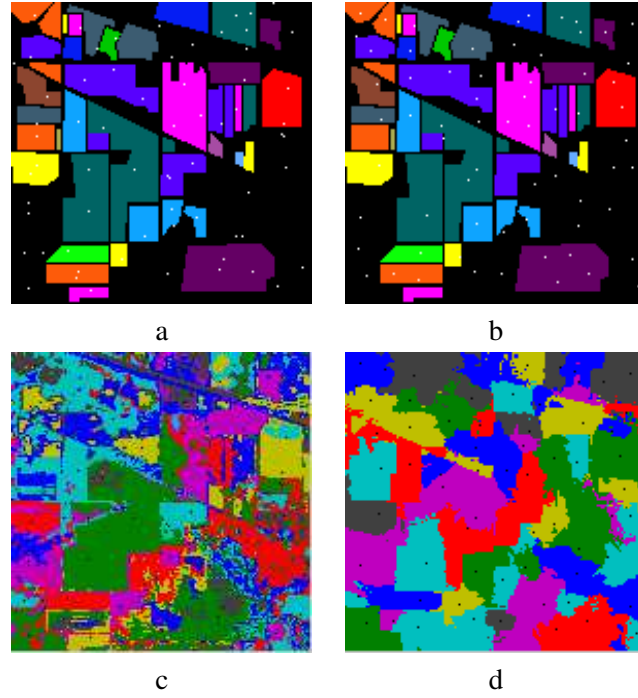


Given two samples whose difference is small, if some feature is numerically enhanced to overcount in the calculation of the difference, this difference can be altered. We suggest to do such a thing with coordinates. Multiplying coordinates by an arbitrary number would make them count more than they did in the differences calculation, so when two samples are spatially close their distance is closer and the way round. Such a number should be decided in terms of the range of the features provided by the spectrometer so the coordinates are overweighed but they do not cause the rest of features be dismissed in the difference. AVIRIS data set ranges its spectral features in [0..255] and its spatial coordinates in [1..145]. We also refer to this as enhanced coordinates.

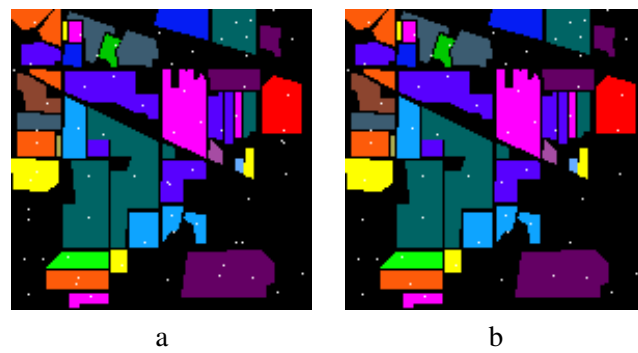
In Figure 4.4 the first row represents the center of the clusters on the groundtruth of the image used. The second row represents the pure cluster result without classifying. This is meant to give an idea on how the samples have been clustered. By using random colors we try to avoid the comparison with the ground truth as this representation is not the result of classification and the clusters do not have class equivalent. This representation is made for the case of the mode seek clustering without over-weighting (b)(c), and the case with coordinates overweighed (b)(d). The main consequence is the homogeneity of the clusters in the space. Notice that the lower right side of the image in Figure 4.4.c is noisy, several clusters are involved in the same area. However, in Figure 4.4.d the same area is covered by 3 clear clusters instead. Therefore, the features were not clear enough to split it into spatial areas and thanks to enhancing the coordinates they are now spatially separated. From the classification point of view the noisy area obtained may be warning about the fact that we are dealing with a heterogenous part of the image. However, our aimed is get nicely distributed centers and we know that connected areas are likely to belong to the same class area. This prevent our training set to have redundant information. Consequently, noticeable differences can be found between Figure 4.4.a-b that represent the centers found over the groundtruth, centers are more distributed in (b) and areas that were missed are now found (look blue area on the top of the image).

When the  $s$  parameter is not properly tuned or such a balance is not possible, see Figure 4.5(a), more samples than needed are selected. The original clustering procedure finds centers of clusters close in the spatial domain. This results in redundant training information. This criterion defines a spatial strategy to select the cluster center with the highest density among those spatially connected. The criterion we suggest consist on merging small nearby clusters based on the class connection principle. Figure 4.5(b) shows the result after applying this criterion. Notice that the two points next each other inside the Corn-no till class (purple) in the center of the image (next the green one), or in the Corn-min till area (orange) below Wheat (light green), in the bottom left part of the image, are reduced to one. The same happens to some other areas. Among them, the sample chosen is the one with highest density which guarantees being representative of a larger amount of samples.

Three different alternatives for including a spatial factor are suggested in this section, they are summarized in Table 4.1. Two of them include spatial information within the selection process, in the clustering algorithm. The other carries out a spatial post-process.



**Figure 4.4:** Representation of 66 clusters and their centers (white dots) over AVIRIS groundtruth (first row) and over the corresponding cluster result (second row). The clustering uses spatial coordinates in both cases. In (a-c) without modification and (b-d) overweighted.



**Figure 4.5:** (a) White dots are the 66 cluster centers resulting from clustering procedure; (b) the previous result after applying the spatial criterion, 58 pixels remain. Both represented as white points over AVIRIS database groundtruth.

	Coordinates	Enhanced coordinates	Discarding neighbourhood
Spatial Information	Includes coordinates in the feature vector	Includes overweighed coordinates in the feature vector	
Post-process			Discards modes in the same neighbourhood

**Table 4.1:** Comparison between the spatial strategies suggested in this chapter.

### 4.1.2 Classification

To select the training data, clustering is carried out using different values of the parameter  $s$  to get different numbers of selected samples, that is, varying the size of the training set. The clustering is performed once per each value of  $s$  to select a certain number of samples that are labeled by the expert. Classification is performed using only the labeled data as training and the rest as test.

#### Distance based classification

A  $k$ -NN with  $k = 1$  classifier is used with the labeled samples as training set. The  $k$ -NN is not an arbitrary choice. Because the clustering procedure used is based on densities calculated on a dissimilarity space, the local maxima correspond to samples which minimize its dissimilarity with a high amount of samples around it. Therefore, these selected samples are highly representative in distance based classifiers. Besides, it is important to highlight that we choose a  $k = 1$ . Since the size of the training set labeled is very small a bigger  $k$  could not be suitable for most of the classes in the image.

#### Label propagation

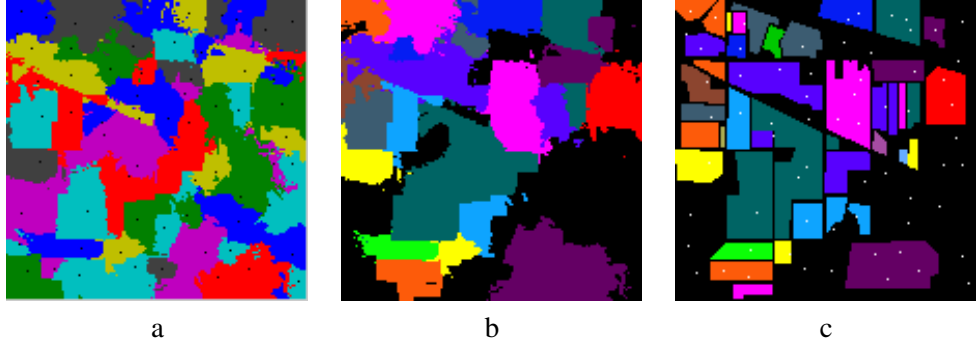
Unsupervised clustering techniques have been proved not to perform segmentation as good as supervised classification methods. This is because the complete lack of information. When performing unsupervised segmentation, each cluster is considered an area or class. Thus, the results are whether over-segmented or not accurate at all. We suggest a hybrid scenario where prior information is not necessary and a non-parametric clustering can be performed to obtain clusters represented by their centers. Then the expert collaborates in labeling the centers which are the most representative samples (the modes). It is a straightforward idea to use the cluster result together with the expert labeling. This idea is based on label propagation, it is a very fast classification that tackle the over segmentation of the unsupervised clustering results by adding the expert collaboration.

The main idea behind label propagation is the cluster assumption. Two samples  $x_i$  and  $x_j$  have a high probability of sharing the same label  $y$  if there is a path between them in regions of significant

density [75]. Many graph-based techniques can be found in literature [20]. To propagate labels using the cluster analysis already performed and according to the main idea of label propagation, we suggest propagating the label of all cluster centers as follows:

Given the set of clusters  $W = \{w_1, \dots, w_T\}$  with centers  $C = \{c_1, \dots, c_T\}$ , the expert assigns labels to the centers  $L = \{(c_1, y_{w_1}), \dots, (c_T, y_{w_T})\}$ . Then  $\forall x_i \in W_t, (x_i, y_{W_t})$ .

The idea suggested is simply propagating labels using the cluster analysis already performed and the labels given by the expert for each center. Figure 4.6(a) shows the cluster result with 66 selected samples and Figure 4.6(b) the classification result when using semisupervised clustering. Note that although the unsupervised clustering is useless on its own, the expert collaboration labeling only 66 points can give an overview of the image (showed on Figure 4.6(c)). Results will be commented in later sections but it is very important to highlight that 66 samples is only the 0.3% of the data set. The 66 samples contain known classes and background. The expert can specify whether a sample belongs to a class or if it is unknown. Thus, we consider all unknown sample as one big background class.



**Figure 4.6:** (a) 66 Clusters represented as an image using random colors; (b) classification result when labels given for the modes are propagated to the rest of the cluster; (c) groundtruth with the modes found marked on white.

### Extracting class frontier information

As presented, the selection scheme is not useful for classifiers that are not based on distances. Given representative centers would not be helpful to find frontiers, as it is necessary, for example, for a SVM. When looking for frontiers it is interesting to detect the separation between clusters and not their centers. Besides, one single point is not enough to represent the shape of the data in the feature space. However, to import the idea of the training set selection to a SVM classifier, we would not like to increase the amount of labeled data. To this end, we bring here the labeling propagation idea used in the semi-supervised clustering. The distribution of the data in the cluster is unknown but we can localize where most of the data is placed.

For a sample  $x_j$ ,  $x_j \in w_t$ . Distances  $D_n = \{d_1, \dots, d_N\}$ , where  $d_j = \text{distance}(c_{w_t}, x_j)$ , and  $sd(D_n)$  the standard deviation of  $D_n$ .

We assign the label  $y_{w_t}$  of the  $c_{w_t}$  to the sample  $x_j$ :

$(x_j, y_{w_t})$  if  $\text{mean}(D_i) + sd(D_i) \leq d_j \leq \text{mean}(D_i) + 2 * sd(D_i)$ .

Calculating the mean and the standard deviation of the distances all samples to the center of the cluster they belongs to is not precise but gives an estimate. Then we could consider the possibility of propagating the label to the whole cluster or also doing it to all the data included in the sphere created taking as a limit  $\text{mean}(D_i) + 2 \times sd(D_i)$  (see Figure 4.7). There are two reasons for discarding these options.

- Propagating the label of the center to all data points in the cluster increases the errors introduced by label propagation since the further a data point is from its center the more possibilities exist that they do not share the same label, according to the cluster assumption.
- We aimed to use a SVM as classifier and training is the most expensive step. Increasing considerably the training data has an undesired effect on the computation time. Therefore, knowing that most of the data is placed around the mean and within the standard deviation, taking data around these two measures seems a good trade-off.

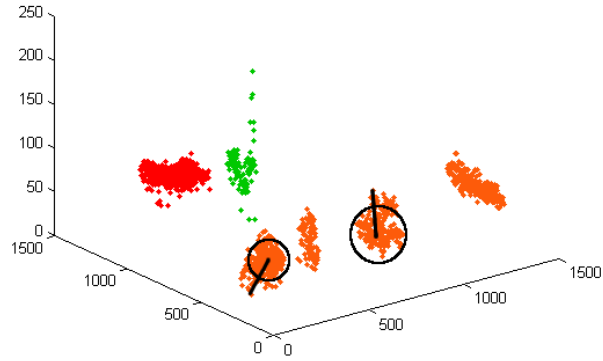
Thus we decided to create a spherical ring ( $\text{mean}(D_i) + sd(D_i) \leq d_j \leq \text{mean}(D_i) + 2 \times sd(D_i)$ ) that can still model a considerable amount of data around the mode. Although we ignore the exact shape of the classes and we assume a sphere, we extract information about where the frontier of the classes are by modeling the samples around the centers. This is an important information needed for classifiers like SVM.

Note that the amount of samples selected is higher than the number of modes, but only the modes are labeled. Consequently, the size of labeled samples stays, although the real size of the training set increases. With this bigger set we can train a SVM and use it to classify the pool of unlabeled samples remaining. It is important not to forget that here the error made by label propagation is introduced in the training set and has to be taken into account when giving the result of this technique.

To get an idea of how the proposal behaves, two standard data sets have been used to test it. All of them are free available [6]. A summary of their properties is shown in Table 4.2. Note that we chose data sets with fairly high dimensionality and a number of unbalanced classes in order to approximate the real case of the hyper-spectral data sets we usually deal with. Notice that because the clustering is distance based, all data has been standardized so that all dimensions are comparable.

In Figure 4.8 the classification error is presented for three different classification strategies. A random selection of the training using a  $k$ -NN classifier (random), a selection of the training set by clustering and a classification using a  $k$ -NN classifier again, and the same selection of the training extended with label propagation and a SVM classifier.

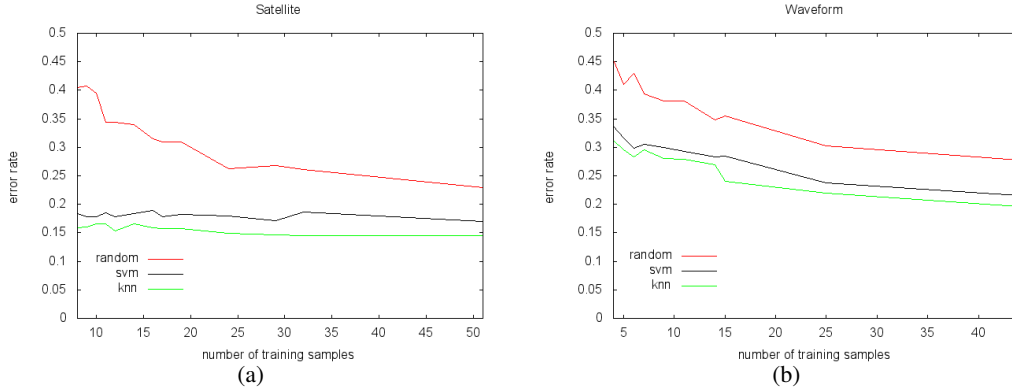
Note that the random case is supervised, has information about the number of classes present and selects a given number of examples proportional to the a priori probability for each class. Despite this, the semi-supervised case outperforms it in the two cases. The clustering method selects the samples to be labeled and samples from all classes are selected even when classes are



**Figure 4.7:** Training selection example for extension of the scheme to SVM necessities. Two modes are highlighted with a point, from the mode a line is drawn to the furthest sample within the cluster and a circle marks the distance in which samples will be selected to extend the label of the mode.

Dataset	NC	S	MaxS	MinS	D
Waveform	3	5000	1696	1647	21
Satellite	6	6435	1533	626	36

**Table 4.2:** Properties of each data set used in the experiments: Name, Number of classes present in the data set (NC), Size of the data set in samples (S), Maximum number of samples per class (MaxS), Minimum number of samples per class (MinS) and Dimensionality (D)



**Figure 4.8:** Learning curve of two small standard data sets to validate the extension of the scheme for SVM using label propagation. Classification is presented in terms of error rate versus the size of training data in number of samples selected by the scheme suggested. The error introduced in the label propagation is also encountered and the training data size stands for the number of samples labeled.

unbalanced. The total training data size obviously is larger for the SVM semi-supervised method as it includes the propagated data as well. However, the data introduced is not labeled so it cannot be considered as an increased of the labeled data. Increasing the training data using label propagation can introduce errors in the training set itself. This error is also taken into account in the global error rate given. Despite this, the classifier generally benefits from a better description of the data improving the result over the supervised random selection classification. Besides, the description is also possible because the label propagation is made from the center of the cluster and that guarantees the highly representation of the label for the samples around it. Even when having prior knowledge about classes, when samples are selected at random, these can lie everywhere in the space and their significance is not guaranteed.

## 4.2 Experimental results

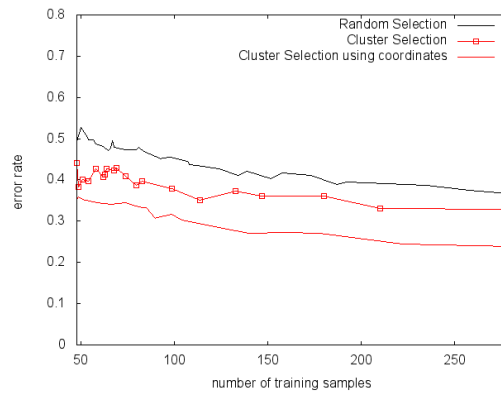
Remember that, to test the method, clustering is performed using different values of the parameter  $s$  to select different numbers samples (the training set size). The clustering is performed once per each value of  $s$  and, as a consequence of the  $s$  value, a number of modes is found. The expert labels these selected samples (modes) and classification is performed using the labeled data as training and the rest as test. Plots are represented in terms of error rate against number of labeled samples and represent the improvement of the scheme with different amount of labeled data (learning curve).

The classification is performed for the known classes. Modes will be found in the unknown area and we assume that the expert dismisses those areas from which they ignore the class. The performances presented in this section are calculated on the known classes.

The datasets are used in their reduced version of 10 spectral bands, using the band selection method. Regarding the parameters related to the spatial criteria suggested before, the neighbourhood in which a center should be dismissed depends on the size of the image and the size of the classes. A big image with big classes would need a bigger neighbourhood than an image with small areas or classes because a big neighbourhood may skip small classes in between. In the case of AVIRIS the size of the smaller class contained is 20 pixels. That is why a neighbourhood of  $9 \times 9$  was chosen. For the rest of datasets we chose  $20 \times 20$  as their classes are not as unbalanced as AVIRIS. The selection method with the different improvements are compared with random selection. Segmentation and per class results are also studied.

### 4.2.1 Influence of the spatial information

The first improvement suggested is adding spatial information to the feature vector of each sample by including the spatial coordinates of the pixel. In Figure 4.9 the learning curves of six different configurations are plotted. First the classification training with random selected samples it is shown, then the selection suggested in this chapter including and not including the coordinates. When coordinates are not used in the clustering, the clusters miss the spatial information so the modes obtained are not describing the classes (as seen in Figure 4.2) and the results, although slightly better than the random selection of the training, are not impressive. When the coordinates are used the clouds of samples in the feature space, are also spatially distributed and it is easier for the clustering to provide a better description of the data. Consequently, the result of the  $k$ -NN classification outperforms considerably the random selection result.



**Figure 4.9:** Learning curve of the  $k$ -NN classifier in terms of error rate when increasing the size of training data in number of samples selected by the suggested scheme showing the impact of including the spatial coordinates as features.



### 4.2.2 Classification by active learning

We have seen that it is convenient to include the coordinates as features in the clustering, they will be in the feature vector for clustering selection from this point on. Notice that this is added for the selection process but in the classification the coordinates do not take part of the feature vector, for direct comparison with other methods in literature. The feature vector considered for classification in this experiments consist of ten spectral features.

The results of the following improvements are presented in Figure 4.10 for three datasets (AVIRIS, CHRIS-PROBA and HYMAP). In all cases, using more training data increases the performance of the classifier but not all methods result in the same performance gain. The distance between random (random  $k$ -NN) and the rest of alternatives is noticeable, it proves that an analysis of the data is preferable to a random pick. Using this method there is a selection that aims at providing suitable and representative data as training. Notice that the scheme without any improvement (selection  $k$ -NN), although it is better than random, can be improved because it still includes redundant training data. We compare the results for the same amount of labeled data, the horizontal axis of the plot corresponds to the size of the training set.

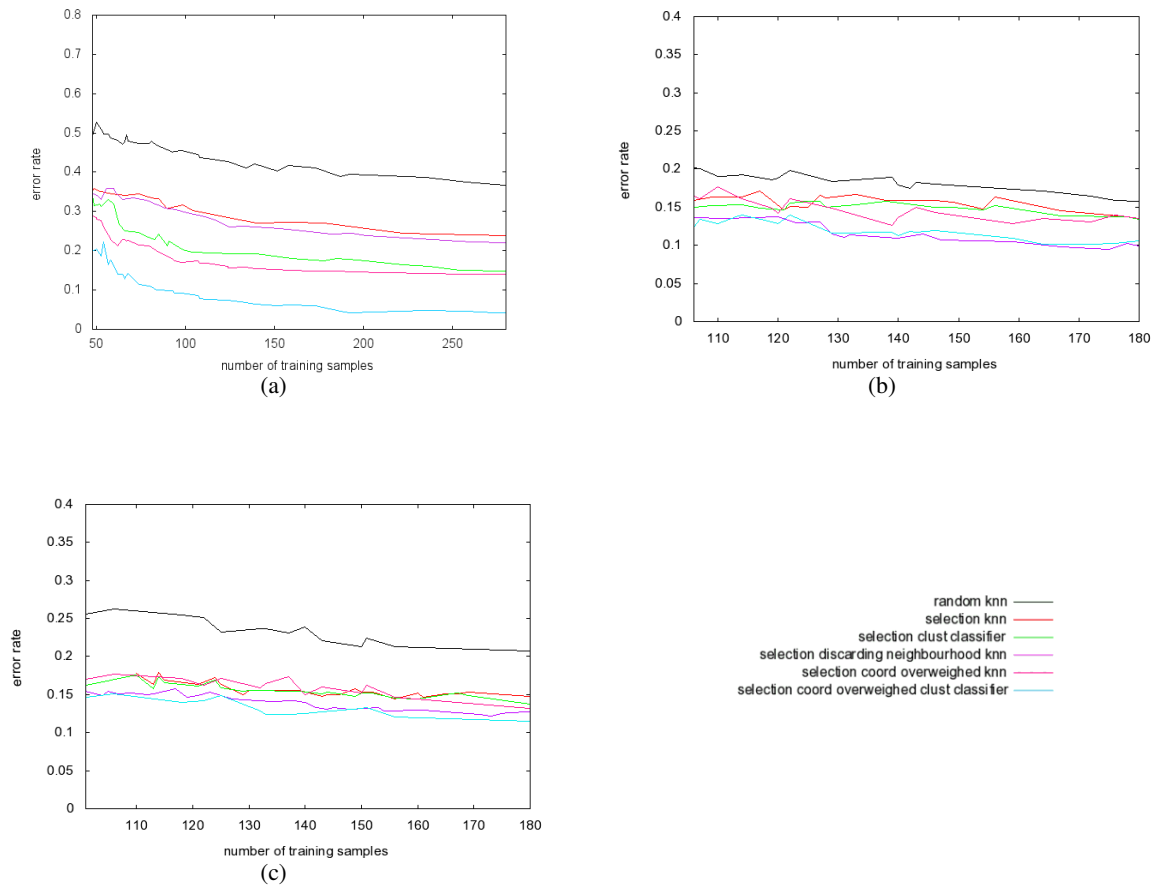
By either discarding centers in the same neighbourhood (selection discarding neighbourhood knn) or enhancing the coordinates role (selection coord overweighed knn) the results outperforms the original scheme. Between the two alternatives, enhancing coordinates in the difference calculation gets a better improvement. It is interesting to point out that it is worthless to try to discard modes in a neighbourhood when coordinates have already been enhanced. Notice that the enhancement of the coordinates force the clustering to cluster together pixels spatially connected. Thus, when trying to spatially discard modes, none is found because they are all included in the same cluster and only one mode is representing them.

Notice also that the results from the basic strategy and the one that discards modes in the same neighbourhood stay together until the number of centers grows. This happens because when including the spatial information (coordinates within the feature vector), the probabilities of finding two modes that are spatially close is very low when the number of clusters is small. The improvement consist in discarding centers over the result of the non-improved. If no centers are found nearby, no center is discarded and the set with and without improvement is the same. If the training set is the same, the result is as good. On the other had, when the number of clusters grows some centers can appear spatially connected. At this moment the strategy of discarding them makes a difference. For the same number of training samples, the set obtained with improvement is equal to the set without improvement less the centers spatially connected plus other new found ones that are neither spatially connected. This mean that for the same amount of label data, the improved method includes a higher number of non-redundant centers which improves the result of the classification.

When the improvement is included in the clustering procedure (overweighed coordinates), the result of the clustering provides a different set of clusters not comparable with the other two. This is because enhancing the role of the coordinates force the clusters to split or merge. This result do not contain spatially connected centers.

Very important differences are found between the two classifying methods. The cluster-based classification (semi-supervised clustering) outperforms the results obtained by training based clas-

sification ( $k$ -NN). When no improvement is included in the clustering the learning curve of the training based classification (red) stays far above from the semi-supervised clustering result (green). The same occurs when overweighed coordinates are used as improvement (magenta), one can find much below the learning curve obtained by cluster-based classification (blue). This stays for all databases tested.



**Figure 4.10:** Learning curve of classification in terms of error rate versus the size of training data in number of samples selected by the scheme suggested with the two improving alternatives compared with the usual random pick. Classification with semi-supervised clustering is also included. In all cases, features consist of 10 spectral features and when a classification is performed,  $k$ -NN classifier with  $k=1$  is used. The results are shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases.

### Feature independency test

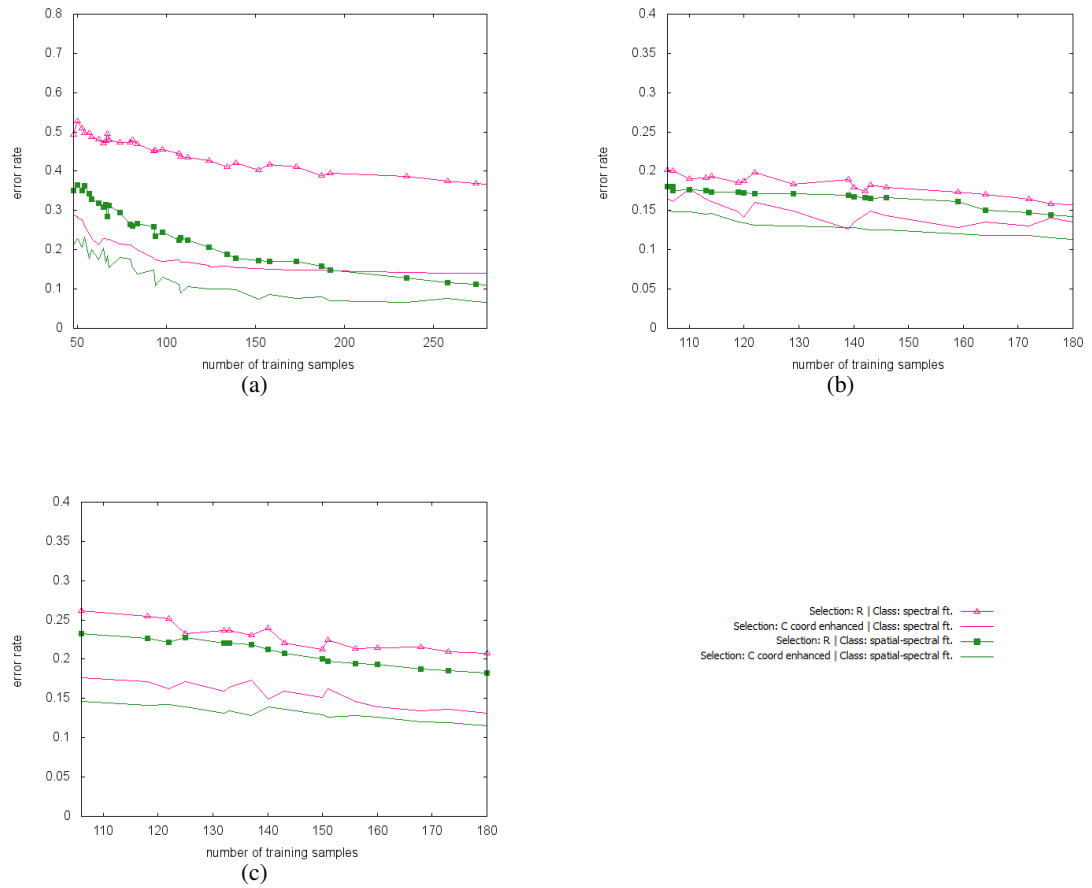
In the description of this technique, we state that the benefits of the selection are independent of the features used a posteriori for classification. We experimentally prove it here by showing the classification results using the same selection technique and two different type of features for classification. Notice that the selection is always performed in the same way and the features vary when classification is to be performed. That can be summarized in the following modified scheme:

1. Band selection for reducing the data set to a given dimension (usually 10).
2. Extending the data by adding the spatial coordinates.
3. Performing unsupervised analysis to obtain modes.
4. Label the modes found.
5. Train a classifier with the samples selected. Here the features can be the ten spectral features of step 1 (coordinates are now dismissed) or the features can be changed to other type of features like the spatial-spectral features defined in Chapter 3.
6. Perform classification and get segmentation result.

Figure 4.11 show the classification results when the training is picked at random and when the selection technique is used. In both cases, two different type of features are used for classification. The comparative between the random selection and the selection method when the spectral features are used was showed previously in this chapter. It is included again in the plots for the three datasets. Observe that the improvement when substituting random selection by our technique is still obtained when the features are changed. Besides, the difference between type of feature is also present. Spatial-spectral features outperform spectral features. The advantage is less noticeable for CHRIS-PROBA and HYMAP. Note that the error is low and there is few room for improvement. This is due to the limits between classes which are inaccurate. Remember that the  $k$ -NN classifier is used with  $k = 1$  and the maximum number of training samples in this plots is 180 which for CHRIS-PROBA and HYMAP represents the 1% and of data 0.2% respectively for each dataset so the possibility of over-fitting is discarded.

### Class frontier information results

When the training set is scarce, quadratic classifiers tend to fail and distance based classifiers are more effective. For quadratic classifiers, sufficient training pixels for each spectral class must be available to allow reasonable estimates of the class conditional mean vector and covariance matrix. For an  $N$  dimensional multi-spectral space Swain et al [95] recommend, as a practical minimum, to use  $10N$  training pixels per spectral class, with as many as  $100N$  per class if possible. Up to now results have been presented for  $k$ -NN classification with  $k = 1$  or semi-supervised clustering. To allow other classifiers to benefit from the strategy we suggest to use label propagation as explained



**Figure 4.11:** Learning curve of classification in terms of error rate versus the size of training data in number of samples. Random pick and selection technique of the training data are used. For the two of them, after selecting the training in the same way, two different type of features are used for classification: 10 spectral features and spatial-spectral features. In both cases, the classification is performed using a k-NN classifier with  $k=1$ . This is shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases.

in Section 4.1.2. The error of the label propagation is included in the error rate but data is described in the space, providing the classifier enough data to be trained.

In Figure 4.12 the best classification alternative with  $k$ -NN classifier and the semi-supervised clustering are compared with the extension of the training selection for SVM classification with a third order polynomial kernel. The result of the last one improves the random selection training strategy but do not outperform the other two strategies, it stays in the same range. The reason is that despite the training is representative, the label propagation sums up error to the SVM classification error. For future, it can be considered work to use better techniques of label propagation.

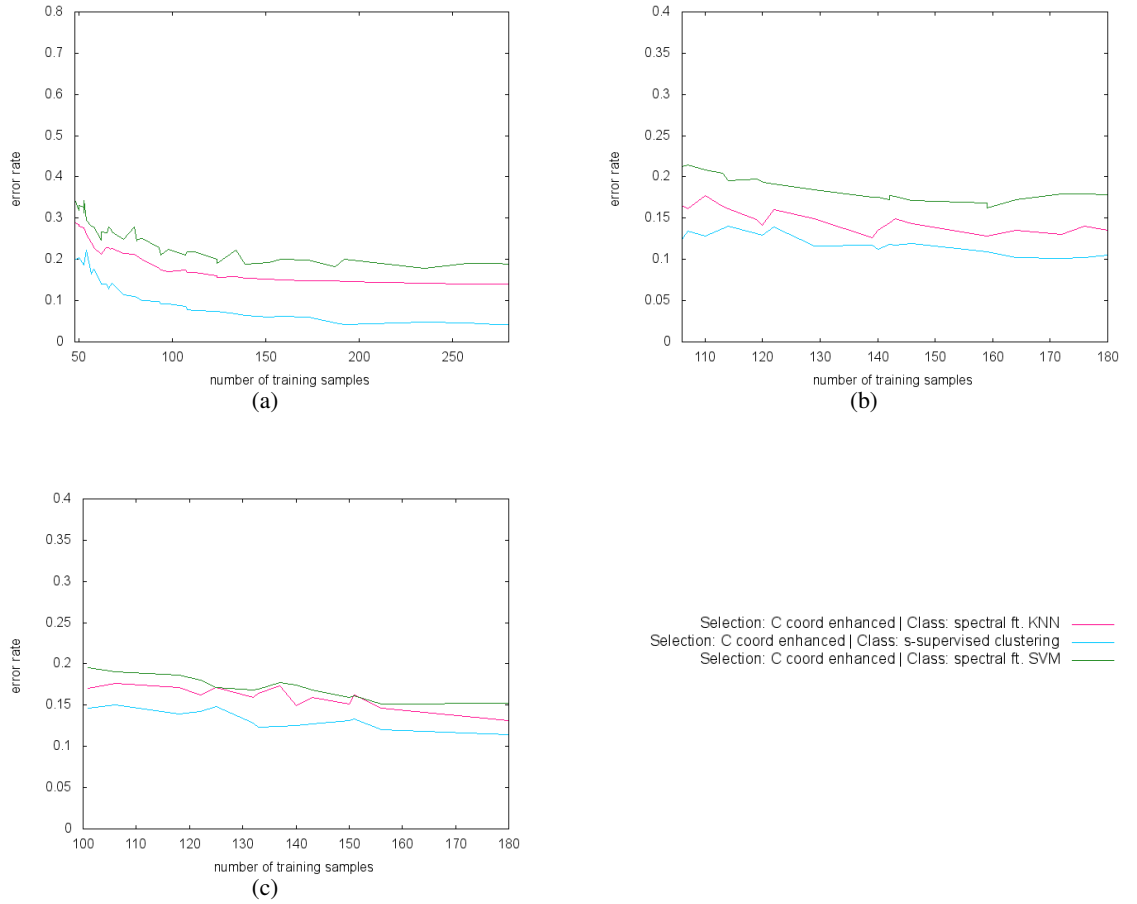
### 4.2.3 Segmentation results

The improvements in error rate (learning curve) are interesting but do not give an overview of what happens in the image in terms of class recognition and segmentation. Up to now we considered that the expert was ignoring the samples selected on unknown areas. From now on, we assume that the expert labels them as a special class that unifies all the unknown area. Observe in Figure 4.13 a case with a reduce number of training samples (70). In Figure 4.13(a) selected pixels come directly from the result of the clustering using  $s = 56$ . In Figure 4.13(b) they result from the clustering using  $s = 44$  and performing a neighbourhood discarding. As a consequence, the same number of selected pixels as before are obtained. Last, in Figure 4.13(c) the selected pixels are found using clustering with  $s = 91$ , but in this case, coordinates were overweighed.

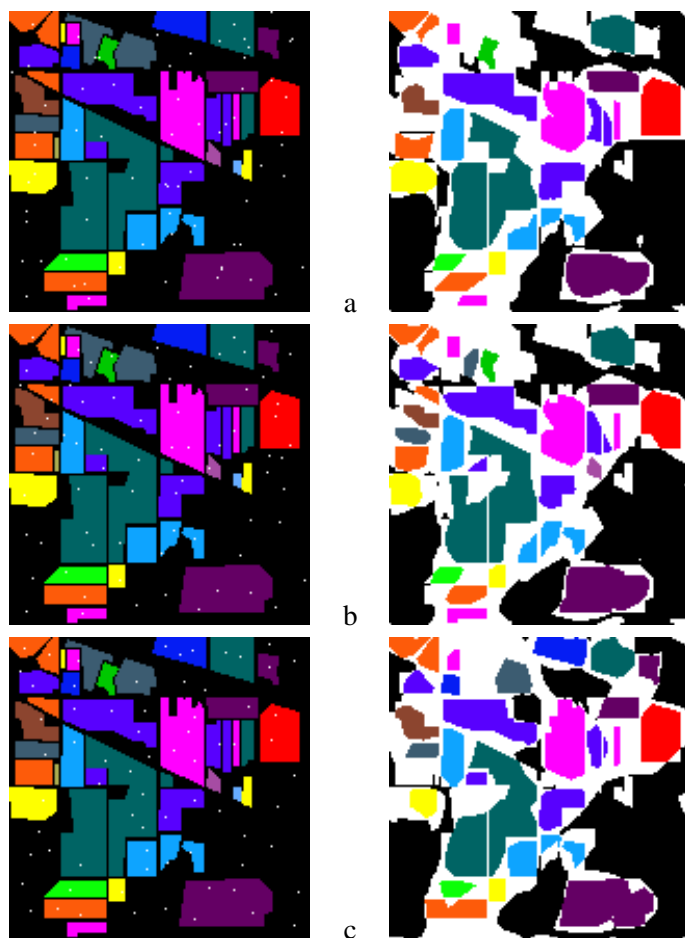
For getting the same amount of selected samples one should use a bigger  $s$  when discarding the neighbourhood to force the clustering to provide more clusters centers and then discard the redundant ones. For the case in which coordinates are overweighed, enhancing the role of the coordinates in the distance calculation make clusters split when samples are spatially away and that provides a larger number of clusters. That is why to get the same number of clusters as the other two alternatives, a bigger  $s$  is needed. Remember that the bigger  $s$  the smaller number of clusters.

To see how this is translated into classification results look the second row of Figure 4.13, the corresponding results can be seen (misclassified pixels are presented in white). Notice that only when coordinates are overweighed the blue area in the top of the image is selected and included for training whereas the rest keeps more or less the same.

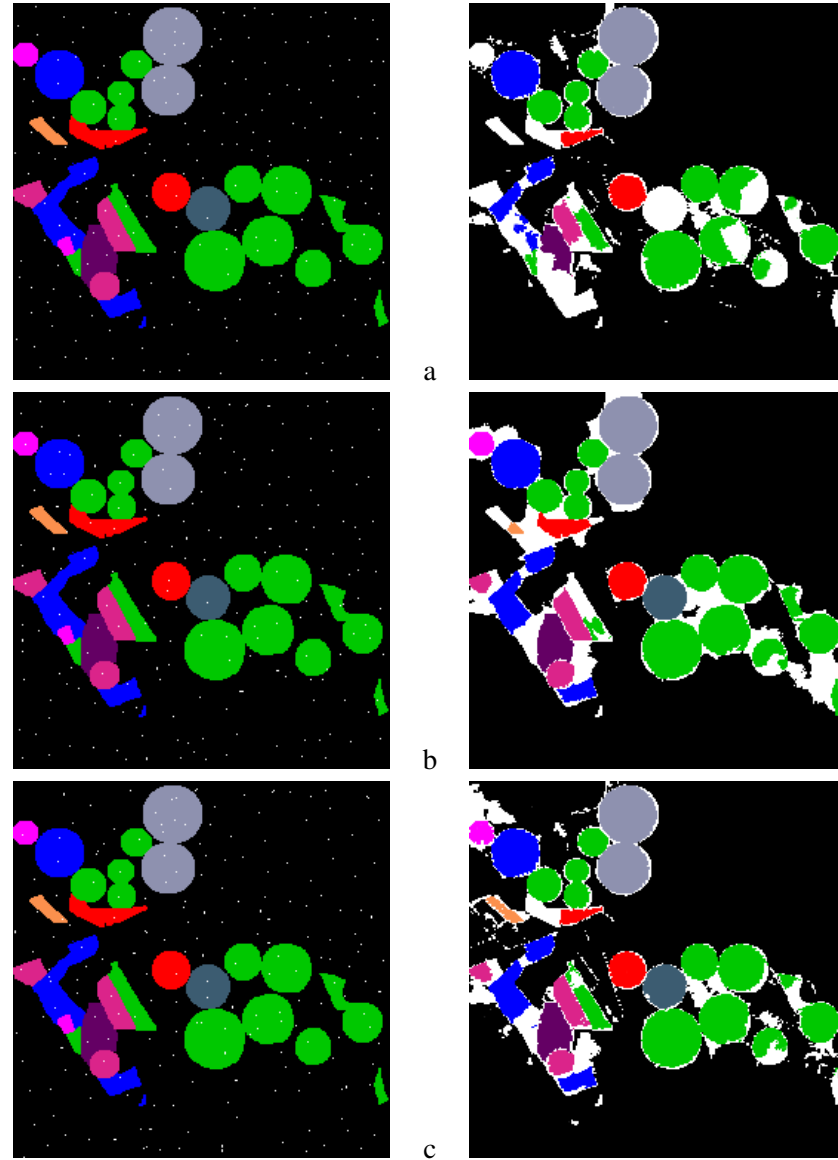
AVIRIS dataset has 21025 samples, that is, 70 samples represent the 0.33% of the data. Let's consider the 2% and the best of our improvements here (coordinates overweighed and cluster classification). See results in the corresponding image from Table 4.3. Also notice that the black area is no longer the background for the classes, it has also been considered as a class, so this is a 17-class segmentation-classification problem. Observe the left top part of the image where the selection manages to detect all of them although the classes are lying one next to each other and their size is not big. The best result is obtained using 4% of the data. The overall accuracy error rate is 0.116 and the most relevant error is the lost of very small classes that can not be found by the clustering.



**Figure 4.12:** Learning curve of classification in terms of error rate versus the size of training data in number of samples selected by the scheme suggested. In all cases, features consist of 10 spectral features but the classification is performed using three different classification algorithms. k-NN classifier with  $k=1$ , SVM with label propagation and semi-supervised clustering. The results are shown for (a) AVIRIS (b) CHRIS-PROBA and (c) HYMAP databases.

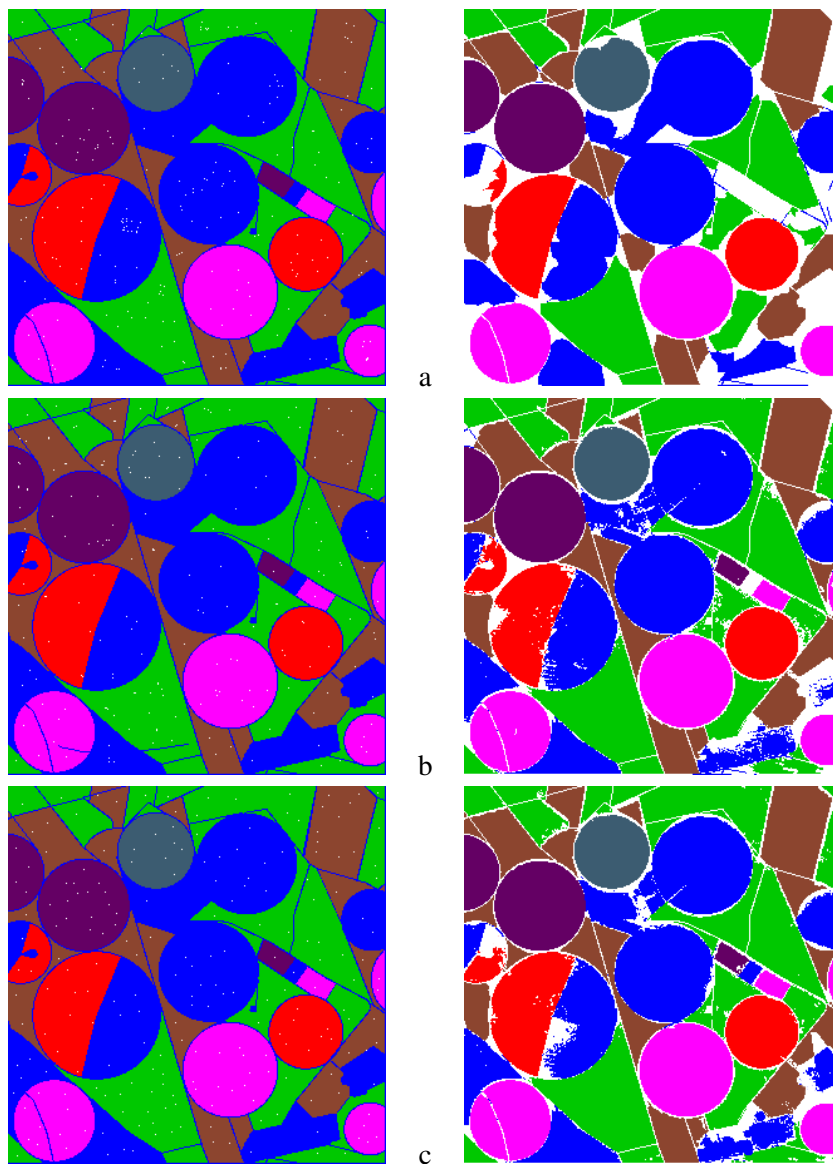


**Figure 4.13:** Representation of the 70 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The right column corresponds to the classification results for each case on the left respectively. The error, misclassified pixels, is represented in white. For AVIRIS dataset.



**Figure 4.14:** Representation of the 220 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The images on the right corresponds to the classification results for each case represented on the left image, misclassified pixels are represented in white. For CHRIS-PROBA dataset.





**Figure 4.15:** Representation of the 220 pixels labeled selected for training by (a) simply clustering. (b) clustering and discarding those lying in the same neighbourhood. (c) clustering overweighing the coordinates of each sample. The images on the right corresponds to the classification results for each case represented on the left image, misclassified pixels are represented in white. For HYMAP dataset.

### Analysis per class
























When facing the analysis per class, overall accuracy is not representative enough. Because classes are highly unbalanced, an increase in the performance is desired when it represents an improvement for all classes and, in this case, bigger classes have a larger impact in the overall accuracy. In Table 4.3 the accuracy per class in error rate is shown, the results obtained with 2% are already comparable, in terms of per class accuracy, with results obtained in supervised scenarios using 10% per class for training [107] or a fixed number per class (50 samples per class, 15 for small ones) [108]. This last approach favors small classes in comparison with the unsupervised selection method presented here. The number of samples per class used here in the training set is unsupervised, unless the clustering detects the class it will be missed. Besides, since the selection is cluster based on the image, the smaller the class is the worse is classified, so in this case small classes are at a disadvantage. Despite this, the accuracy for very small classes are better than in supervised experiments, stone-steel towers, alfalfa, grass/pasture-mowed have accuracies around 0.03 with only between one and five samples. Other classes usually dismissed because of their size [30][13] wheat, corn and Bldg-Grass-Tree-Drives have errors between 0.005 and 0.07 using only seven, six and nine labeled samples. Because here the background class is used, the percentages of data include this class too. Therefore, is fair to mention that dismissing the background, the experiments of 0.33%, 2% and 4% are equivalent to 0.46%, 2.4% and 4.8% of the non-background data respectively.

Regarding datasets CHRIS-PROBA and HYMAP, results are also shown in Tables 4.4 and 4.5 respectively. Due to the smaller number of classes present in both data sets and the low unbalancing between them the amount of training data can be considerably reduced and still produce satisfying results. It is remarkable that, in Table 4.4, classes like Grass or Pot only need 1 or 2 samples of training to be fine recognized. Notice that we compare three different sizes of training set, this means that for each of them the parameter  $s$  changes. When the parameter  $s$  changes, the clustering algorithm finds new clusters, that is, new centers. Thus, even if some sets may have the same amount of training samples for certain classes, this samples are not the same ones. The improvement in the training set can be given by the increase of the number of training samples but also but the quality of the samples chosen. As for Table 4.5, it is remarkable the performance achieved for the classes Cereals and Alfalfa. HYMAP is specially interesting for this method proposed because possesses a complete groundtruth











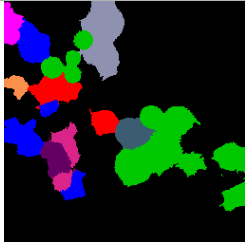
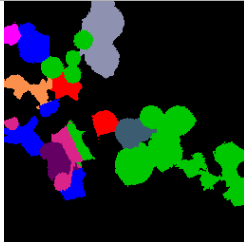
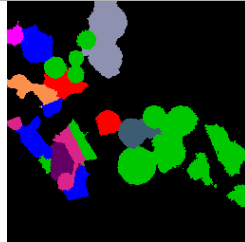

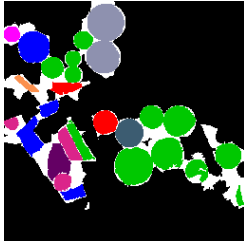

## 4.3 Conclusions

To face the concern of finding a tradeoff between the unavailability of the experts and the necessity of training data, we suggest an unsupervised technique to improve and decrease the amount of labeled data needed. This is useful when no prior knowledge is available and expert collaboration is limited. Thanks to the selection of the training set, only relevant samples are shown to the expert to be labeled. In this sense, expert collaboration is reduced while performance has shown to be raised in comparison with random selection.








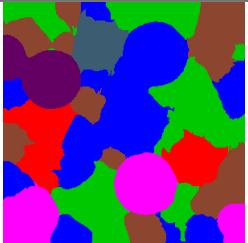
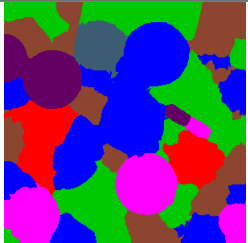
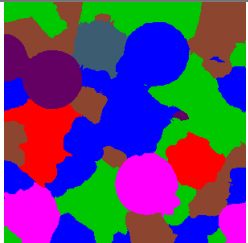
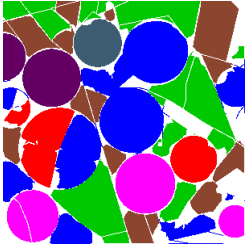
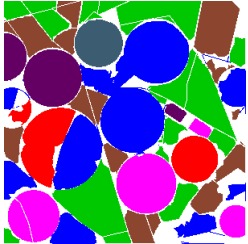
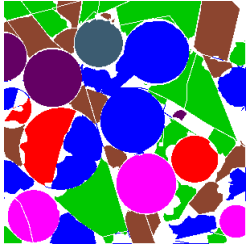
The method is based on an unsupervised study of the data by a clustering technique. Besides, a spatial improvement was suggested to avoid redundant training data. This forced clusters to merge

classes		0.3% of training data		2% of training data		4% of training data	
		training/total	error	training/total	error	training/total	error
Heterogenous background		22/10659	0.432	171/10659	0.262	367/10659	0.193
Stone-steel towers		0/95	1	2/95	0.139	5/95	0.033
Hay-windrowed		2/489	0.004	10/489	0.004	25/489	0.004
Corn-min till		5/834	0.214	18/834	0.076	40/834	0.045
Soybeans-no till		5/968	0.185	25/968	0.060	40/968	0.072
Alfalfa		0/54	1	1/54	0.038	3/54	0.039
Soybeans-clean till		2/614	0.488	15/614	0.066	28/614	0.056
Grass/pasture		3/497	0.105	12/497	0.064	28/497	0.042
Woods		6/1294	0.023	29/1294	0.034	58/1294	0.026
Bldg-Grass-Tree-Drives		3/380	0.021	9/380	0.011	12/380	0.011
Grass/pasture-mowed		0/26	1	1/26	0.040	1/26	0.040
Corn		1/234	0.601	6/234	0.070	10/234	0.049
Oats		0/20	1	0/20	1	0/20	1
Corn-no till		6/1434	0.278	35/1434	0.067	63/1434	0.035
Soybeans-min till		10/2468	0.069	70/2468	0.023	143/2468	0.018
Grass/trees		4/747	0.067	18/747	0.033	34/747	0.042
Wheat		1/212	0.009	7/212	0.005	11/212	0.005
Overall error			0.299		0.156		0.116
kappa			0.685		0.771		0.795
							
							

**Table 4.3:** Accuracy per class for the 17 classes classification of the AVIRIS dataset using semi-supervised clustering classification on 12 features (ten spectral features and two spatial coordinates). For a training set of 0.3%, 2% and 4% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white.

classes		0.2% of training data		0.35% of training data		0.6% of training data	
		training/total	error	training/total	error	training/total	error
Heterogenous background		100/48382	0.0927	177/48382	0.0555	262/48382	0.0659
Pot		4/1050	0.0755	3/1050	0.2026	7/1050	0.0824
Alfalfa		6/2272	0.1377	7/2272	0.2379	18/2272	0.0803
Corn		15/6679	0.2151	23/6679	0.1286	38/6679	0.0839
Garlic in greenhouse		1/340	0.3244	1/340	0.3510	2/340	0.3520
Grass		1/690	0.0842	2/690	0.0436	4/690	0.0495
Onion		3/1250	0.3557	4/1250	0.2044	5/1250	0.2108
Garlic		2/708	0.0071	3/708	0.0141	9/708	0.0171
Sugar cain		1/521	0.5245	2/521	0.5960	2/521	0.118227
Sunflowers		7/3133	0.0691	8/3133	0.0534	23/3133	0.0537
Overall error			0.1127		0.0778		0.0717
kappa			0.7443		0.8154		0.8354
							
							

**Table 4.4:** Accuracy per class for the 10 classes of the CHRIS-PROBA dataset using semi-supervised clustering classification with 12 features (ten spectral features and two spatial coordinates). For a training set of 0.2%, 0.35% and 0.6% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white.

classes		0.5% of training data		2.3% of training data		4.7% of training data	
		training/total	error	training/total	error	training/total	error
Cereals		47/28485	0.2673	74/28485	0.2581	117/28485	0.2512
Barley/sunflower		24/7686	0.0844	38/7686	0.0943	48/7686	0.0869
Vegetables		5/2833	0.0056	10/2833	0.0142	7/2833	0.0187
Corn		38/21831	0.2365	80/21831	0.1751	84/21831	0.1925
Alfalfa		25/9419	0.0811	37/9419	0.0418	49/9419	0.0832
Onions		33/5898	0.0828	24/5898	0.0180	40/5898	0.0580
Fallow land		31/14449	0.1598	43/14449	0.1595	55/14449	0.1548
Overall error			0.1878		0.1628		0.1705
kappa			0.7668		0.7977		0.7878
							
							

**Table 4.5:** Accuracy per class for the 7 classes of the HYMAP dataset using semi-supervised clustering classification with 12 features (ten spectral features and two spatial coordinates). For a training set of 0.5%, 2.3% and 4.7% of the total data (this counts with the background as a class). The last two rows show the segmentation result for each case and the spatial visualization of the error in white.

or split according to the class connection principle. Thus, the training set is representative and free of redundancies.

The selection has shown to be valid for building a classifier even if the features are changed. It was shown that textural-spatial features can also benefit from this selection scheme and achieve same results with less training data. Indeed, results shown outperform results of classification methods in literature that use a random selection of their training set. Moreover, the process does not need large amounts of data since it has been shown that not all spectral bands and not a high number of features were needed in our experiments.

## Conclusions

We had three objectives in this thesis. First improving the state of the art results of pixel classification by including spatial information in the characterization of pixels. Second, achieving the first objective not falling into the curse of dimensionality. The last objective was facing the problem of the scarcity of training data. In practise, these objectives are meant to allow to build specific sensors that reduce costs of production and transmission of the data. Besides decrease the collaboration of the expert to decrease also costs and waiting time.

We achieved the first objective by proposing three pixel characterization methods. The three are spectral-spatial methods based on Gabor filters. These perform spatial-spectral feature extraction in hyperspectral pixel characterization. One does not use inter-channel information (Gabor), whereas the other two use it. For the case of Gabor complex the inter-channel information is obtained by creating complex bands before applying the transform. The opponent features computation does not include the intra-channel information in the Fourier decomposition but combines the responses between channels after the responses are obtained.

The second goal was tackled with a new schema for classification and segmentation of hyperspectral landscape imaging. It starts using an unsupervised method for reducing the dataset, then spatial-spectral characterization is applied to replace the spectral vector traditionally used. Last, it performs per pixel classification providing the direct result, a classification/segmentation map. On the top of them, we faced the third of our objectives designing an unsupervised technique for training selection. This allows to decrease the collaboration with the expert.

It has been experimentally proven that the proposed scheme, with the characterization methods, provided remarkable results in datasets with extreme unbalanced classes. Furthermore, the approach was able to perform using a very limited set of spectral bands, simplifying the representation.

We showed that the spatial information provided an appropriate characterization of the pixels, more than the inter-channel information suggested to be used in other methods. The influence of the different scales in the feature extraction process was studied. We found that, for land-use hyperspectral images, the lower scales provide the best characterization and the addition of the last

scales tends to worsen the classification results. However, if we have to deal with non-homogeneous regions, the use of the medium scales may improve the characterization.

The segmentation experiments show a smooth result, the inner part of the regions was always remarkably homogeneous without needing any spatial post-regularization. Error was mainly found in the borders between classes, due to the transitions between different classes in the image plane

The unsupervised training selection technique was an innovative idea that found a tradeoff between the unavailability of the experts and the necessity of training data. The selection informs about which samples are more suitable for training. That selected set showed to be valid for building a classifier based on distances but also in frontiers, using the idea of label propagation. Besides the classification could be performed with different type of features.

Therefore, our training selection technique could be added to the classification-segmentation scheme suggested, before building the classifier. Indeed, the scheme, once added the selection of the training, outperformed the results of classification methods in literature. Notice that all this process is done using only a few selected spectral bands.

## 5.1 Future work

This thesis opens two main directions for further work: first, the practical application of the developed methodologies; second, new scientific contributions in the same or similar research lines.

The straight forward application to the methodologies here suggested involve the automatic generation of large scale maps. Although we have only dealt with land-use images, the creation of maps of chlorophyll or vegetation indexes can also be faced using these techniques. Several fields of application can be found related to Earth observation. For example, the analysis of hydrological resources and desertification or studying the impact of forest fires are also areas where our proposals can be applied.

As for research lines, we can enumerate the following interesting tasks:

- Continuing with our proposal of extracting spectral-spatial features before the classification process, there exists a large variety of textural methods that could be studied. Although we have chosen the use of Gabor filters, techniques based on mathematical morphology, co-occurrence matrices, other filters, or even combination of these, could be taken into account.
- Regarding the prototype selection to build the training set, a study of different active learning techniques could be interesting. Furthermore, due to the great variety of clustering methods and their different properties, an analysis of how other clustering techniques can fit the methodology suggested could provide interesting findings.
- Although we have been working with band selection techniques, we can find situations where all the spectral bands will be always available. In such cases, the goal is not to reduce the amount of transmitted data nor building a specific sensor. Then, the use of feature extraction techniques, like Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) or Independent Component Analysis (ICA), may be analyzed.



# Chapter 6

## Sinopsis de la tesis

This chapter fulfills a requirement of the Spanish PhD regulation RD 99/2011, which states the criteria to obtain the International mention in the PhD title. In particular, it specifies that part of the thesis has to be written in a different language to the official ones, which are Spanish and Valenciá, but at least the abstract and the conclusions have to be given in one of those official languages as well. Thus, the aim of the following sections is mainly to summarize the previous chapters that have been reported in English, including motivation and general objectives, contributions and conclusions and future work lines.

Este capítulo da respuesta al artículo 15 de la normativa de los estudios de doctorado, regulados por el RD 99/2011, que establece los criterios aplicables para la obtención de la Mención Internacional en el título de Doctorado. Éste establece que parte de la tesis doctoral debe redactarse en una lengua no oficial que se utilice de forma habitual para la comunicación científica en el campo de conocimiento en que se enmarque la citada tesis, y que en cualquier caso, el resumen y las conclusiones se adjunten también en una de las lenguas oficiales. Al haberse redactado la tesis en inglés, se presenta a continuación una visión de conjunto de todo el trabajo realizado en esta tesis doctoral, incluyendo la motivación, objetivos, contribuciones, conclusiones y líneas de trabajo futuras.

### 6.1 Motivación

Comúnmente se llama luz a la parte del espectro electromagnético que puede ser percibida por el ojo humano. Sin embargo, el espectro abarca desde las ondas de radio, pasando por las microondas, los rayos infrarrojos, la luz visible, la radiación ultravioleta, los rayos X y finalmente, los rayos gamma. Todas estas emisiones son producidas por la misma fuente de luz que nos ilumina y se diferencian por lo que se llama la longitud de onda.

Las cámaras que todos conocemos son sensores que capturan la respuesta de los escenarios a la luz visible. Es decir, las imágenes capturadas por esos sensores sólo contienen la respuesta del

espectro electromagnético a la luz visible. El color es la respuesta de los objetos y sustancias a la luz. Sin embargo éstos responden al espectro electromagnético completo de una manera diferente. Si pudiéramos adquirir la respuesta al espectro electromagnético completo de todas las sustancias u objetos de una escena obtendríamos más información que el color. En tal caso dos objetos del mismo color pero de una sustancia diferente responden a la luz de la misma manera (tienen el mismo color) pero pueden tener respuestas muy diferentes en el resto del espectro electromagnético. Un sensor hiperespectral es aquel capaz de capturar la respuesta de un escenario a una parte más amplia del espectro electromagnético. Como este es continuo, dicho sensor discretiza la respuesta obteniendo una respuesta para un rango de longitudes de onda. El número de rangos que el sensor es capaz de hacer es su resolución espectral. A mayor resolución espectral, mayor número de respuestas recogidas. Esto permite un análisis más detallado de la respuesta de objetos o sustancias. Esta respuesta puede ser utilizada para diferenciarlos de forma automatizada.

Cuando un sensor es capaz de recoger esa información de forma matricial para un escenario obtenemos una imagen. Esta imagen tiene tres dimensiones, dos dimensiones espaciales y una tercera espectral. El tamaño de la dimensión espectral es igual a la resolución espectral del sensor con el que fue adquirido. Cada punto de esa imagen (píxel) representa una parte del escenario. Cuando más pequeña es la parte representada por un píxel más detalle aporta la imagen. Esto es la resolución espacial del sensor. Cada vez se construyen sensores con mayor resolución espectral y espacial. Esto aporta mayor detalle. Sin embargo puede generar problemas de computación y de presupuesto. A mejor sensor mayor coste.

Nuestro interés se centra en las imágenes aéreas tomadas por sensores hiperespectrales desde un avión o un satélite. En este caso cada píxel de la imagen es una parte de la superficie y puede ser clasificado según lo que se encuentra en ella. Clasificando cada punto de la imagen se obtiene un mapa de clasificación en el que los puntos forman áreas de la misma categoría. Esto es conocido como segmentación. De tal manera, en esta tesis pretendemos crear mapas de clasificación de la superficie terrestre. Para crear un sistema automático de creación de estos mapas, necesitamos datos conocidos para entrenar nuestro sistema y que este se enfrente después a datos desconocidos. Debe considerarse que, cuando las imágenes representan una superficie de kilómetros, adquirir datos supone que un equipo de expertos se mueva por toda la superficie del escenario capturado.

## 6.2 Objetivos

Cada vez se construyen sensores con mayor resolución espectral y espacial. Esto aporta mayor detalle de la superficie a explorar. Sin embargo, esto también puede generar problemas de computación y de presupuesto, dado que a mejor sensor mayor coste. En esta tesis nos enfrentamos al problema de reducción de información para abaratar costes. Igualmente, como nuestro interés son las imágenes aéreas, cuando el sensor se encuentra en un satélite, minimizar la cantidad de datos necesarios también reduce considerablemente el tiempo de transferencia. Otro problema a tener en cuenta es la adquisición de datos para el entrenamiento de nuestros sistemas. Cuando las imágenes representan una superficie de kilómetros adquirir datos supone que un equipo de expertos se mueva

por toda superficie del escenario capturado. En tal caso reducir la cantidad de datos conocidos es importante. Este problema también es tratado en esta tesis doctoral.

En la literatura se pueden encontrar muchos métodos para la clasificación y segmentación de imágenes aéreas [106] [107] [64] [10]. Sin embargo, estas técnicas utilizan toda la información espectral y algunas incorporan información espacial de forma que la cantidad de información es considerable y las propuestas poco escalables. Nuestro objetivo es proponer un esquema de clasificación alternativo que necesite menos información. Nuestros objetivos se resumen en:

1. Mejorar el estado del arte en la clasificación automática de imágenes hiperespectrales.
2. Evitar el uso de grandes cantidades de información para prevenir problemas con la dimensionalidad.
3. Diseñar técnicas que permitan disminuir la cantidad de datos de entrenamiento mientras se mantiene el rendimiento.

## 6.3 Contribuciones

De acuerdo a los objetivos mencionados, las principales contribuciones se pueden estructurar en los siguientes apartados.

### 6.3.1 Esquema alternativo para la creación de mapas de superficie terrestre

Las técnicas existentes clasifican los píxeles de la imagen utilizando la información espectral. Al resultado que esto genera se le aplican correcciones espaciales. También pueden encontrarse métodos que extienden la información espectral con información espacial antes de la clasificación.

La caracterización espectral-espacial propuesta es realizada mediante filtros de Gabor. Este tipo de filtros son capaces de analizar simultáneamente orientación y frecuencia [81]. Cada filtro tiene dos dimensiones: una orientación y un rango de frecuencias espectrales. Al rango de frecuencias nos referiremos como escala. Este tipo de filtros nos permiten lograr un análisis conjunto de la frecuencia y del espacio [39].

En esta tesis se sugiere un esquema en el que primero se caracterizan los píxeles de forma espacial y espectral. Estas nuevas características sustituyen a las espectrales de forma que el resultado de la clasificación es directamente el mapa de clasificación.

Además sugerimos incorporar un pre-proceso no supervisado (automático) de selección de bandas de forma que la creación de las características se realiza sobre una cantidad reducida de información.

### 6.3.2 Reducción de la dimensionalidad mediante el análisis de la información

Los filtros de Gabor se caracterizan por una escala de frecuencia y una orientación. Una forma de reducir la información utilizada es realizar un estudio de las escalas para comprobar si el desempeño del método mejora cuando se utilizan todas las escalas o si existe un subconjunto de estas que

permite mejorar el resultado. De ser así la cantidad de información se puede reducir utilizando ese subconjunto, esto puede variar dependiendo de la naturaleza de las imágenes. En esta tesis el análisis se realiza para el tipo imágenes que nos ocupan.

### 6.3.3 Selección de los datos de entrenamiento del sistema

En los métodos encontrados en la literatura los datos de entrenamientos son elegidos al azar de entre todos los disponibles. Este escenario es posible cuando tenemos muchos datos conocidos disponibles. Sin embargo en muchas ocasiones nos enfrentamos al caso en el que todos los datos son desconocidos y necesitamos trabajar con un experto o un grupo de ellos que nos proporcionan el conocimiento necesario de un grupo de datos con el que podremos entrenar nuestro sistema. En tal caso, encontramos literatura que sugiere técnicas de aprendizaje interactivo [98][65]. Sin embargo en el caso que nos ocupa interactuar con el experto es costoso y tedioso.

En esta tesis sugerimos un método de selección de los datos para que estos puedan ser proporcionados al experto de una vez sin necesidad de interactuar repetidamente. El método utiliza una técnica de análisis de datos no supervisado (clustering) mediante la cual los datos son agrupados según cercanía en el espacio de características. De cada grupo encontrado existe un representante. El conjunto de representantes son los datos seleccionados para ser analizados por el experto y seguidamente utilizados como datos de entrenamiento para el sistema.

### 6.3.4 Difusión del trabajo de investigación

Se ha llevado a cabo un esfuerzo por conseguir la difusión y reconocimiento de la comunidad científica de este trabajo. Los resultados de esta tesis se han dado a conocer progresivamente dentro de la comunidad científica mediante numerosas publicaciones, la lista de las cuales puede encontrarse en la introducción de esta tesis y una copia completa de las mismas en un Anexo. Así mismo el trabajo ha sido reconocido como innovador con un premio a la mejor contribución en una de las conferencias más reconocidas en el campo de trabajo.

Además, las colaboraciones que se han realizado dentro del plan en el que se ha enmarcado esta tesis, han llevado a la utilización de este método en un ámbito diferente como es el de la imagen médica.

## 6.4 Conclusiones

Esta tesis tenía tres objetivos. Primero mejorar el estado del arte de la clasificación de píxeles para imágenes aéreas de superficies terrestres. El segundo era conseguir el primer objetivo reduciendo la utilización de información. El último de los objetivos consistía en proponer una solución para la escasez de datos de entrenamiento y la interacción con el experto. En la práctica estos objetivos se traducen en la capacidad de crear sensores de propósito específico para reducir costes y transmisión de datos.

El primer objetivo se consiguió con la propuesta de un esquema de clasificación donde primero se seleccionan longitudes de onda de la resolución espectral del sensor utilizado para adquirir la

imagen. Después cada píxel se caracteriza de forma espacial-espectral mediante la utilización de un banco de filtros de Gabor. De esta forma la clasificación ya contiene información espacial y no es necesario un post-proceso del resultado. Este tipo de filtros permite un estudio por escalas permitiendo descartar el uso de aquellas escalas que, para determinado tipo de imágenes, proporcionan características de los píxeles que no mejoran el rendimiento de la clasificación. Esto fue utilizado para lograr el segundo de los objetivos.

El tercer objetivo intentaba solucionar una problemática diferente, la interacción con el experto. Se intentaba una menor interacción con el experto y de una manera más directa y concisa. Esto se consiguió mediante el diseño de una técnica no supervisada para la selección de datos de entrenamiento que proporciona, con sólo una interacción, los datos más apropiados al experto para que este los analice.

Los resultados experimentales del esquema muestran resultados notables con imágenes con clases de tamaños muy diferentes (zonas muy grandes y muy pequeñas dentro de la imagen). Es más, esto se consigue haciendo uso de muy pocas bandas espectrales lo que simplifica mucho la necesidad de datos y posibilita la utilización de sensores menos sofisticados. La cantidad de información puede reducirse más si se seleccionan las escalas de frecuencia apropiadas para el tipo de imagen que se vaya a analizar. Imágenes con zonas amplias necesitan pocas escalas de bajas frecuencias. A medida que se incrementa la presencia de áreas más pequeñas o detalles, las escalas de frecuencias medias son necesarias. En cualquier caso escalas pertenecientes a altas frecuencias son en cualquier caso despreciables debido a la alta cantidad de ruido que incluyen.

Es importante destacar que los resultados de segmentación son suaves. Es decir, el conjunto de píxeles clasificados de una misma manera que se encuentran espacialmente conectados, es coherente y crea áreas de clasificación homogéneas. Esto se consigue con el resultado directo de la clasificación sin necesidad de un post-proceso espacial. Los principales errores se localizan en las transiciones entre áreas.

El método no supervisado de selección de datos de entrenamiento es una idea innovadora que encuentra un balance entre la disponibilidad del experto y la necesidad de datos de entrenamiento. La selección informa sobre qué datos son mejores para entrenar el sistema y el experto sólo ha de analizar esos. En la tesis se muestra que el método puede utilizarse con clasificación basada en distancias y en fronteras. Además es igualmente efectivo con diferentes tipos de características.

Finalmente la selección de datos de entrenamientos puede ser añadida al esquema inicialmente propuesto antes de la creación del clasificador que generará el mapa final. De esta manera los resultados obtenidos mejoran otros resultados encontrados en la literatura a pesar de utilizar una cantidad de información muy reducida.

## 6.5 Líneas de trabajo futuras

Esta tesis abre dos vías principales de continuación: primero, la aplicación práctica de las metodologías sugeridas; segundo, nuevas aportaciones científicas en una línea similar. La aplicación directa de los métodos propuestos es la creación de mapas de gran escala. Aunque en este caso nos hemos limitado a mostrar resultados de clasificación según el tipo de terreno, la creación de mapas de

clorofila o índices de vegetación pueden ser abordadas utilizando las mismas técnicas. En relación con la observación de la Tierra existen diferentes aplicaciones donde estas técnicas son de utilidad: análisis de recursos hidrológicos, estudio de la desertización o estudio del impacto de incendios forestales.

Respecto a las líneas de investigación que podrían seguirse, el trabajo podría extenderse mediante:

- Estudio de otras técnicas de caracterización para el esquema de clasificación propuesto. En este caso se han utilizado filtros de Gabor pero se pueden encontrar multitud de métodos de análisis de texturas para los que es interesante estudiar su portabilidad al uso con imágenes hiperspectrales.
- Estudio del comportamiento de otras técnicas de análisis no supervisado para la selección de datos de entrenamiento y su comparación con los métodos de aprendizaje activo.
- Investigación de otros métodos de selección o extracción de características en el pre-proceso del esquema de clasificación como Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) o Independent Component Analysis (ICA).

# Bibliography

- [1] A., B., AND L., S. Fuzzy rule-based classification of remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing* 40, 2 (2002), 362–374.
- [2] A., P., P., M., R., P., AND J., P. A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on* 42, 3 (march 2004), 650 – 663.
- [3] A.FARAG, R.MOHAMED, AND A.EL-BAZ. A unified framework for map estimation in remote sensing segmentation. *IEEE Trans. on Geoscience & Remote Sensing* 43 (2005), 1617–1634.
- [4] A.F.H., G., G., V., J.E., S., AND B.N., R. Imaging spectrometry for earth remote sensing. *Science* 228, 4704 (1985), 1147–1153.
- [5] ARYA, S., MOUNT, D. M., NETANYAHU, N. S., SILVERMAN, R., AND WU, A. Y. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM* 45, 6 (1998), 891–923.
- [6] ASUNCION, A., AND NEWMAN, D. UCI machine learning repository, 2007.
- [7] BALI, N., AND MOHAMMAD-DJAFARI, A. Bayesian approach with hidden markov modeling and mean field approximation for hyperspectral data analysis. *IEEE Trans. on Image Processing* 17, 2 (2008), 217–225.
- [8] BAU, T., SARKAR, S., AND HEALEY, G. Hyperspectral region classification using three-dimensional gabor filterbank. *IEEE Trans. on GRS* 48, 9 (2010), 3457–441.
- [9] BELLMANN, R., AND CORPORATION, R. *Dynamic Programming-Code*. A Rand Corporation Research Study Series. Princeton University Press, 1957.
- [10] BENEDIKTSSON, J., J.A., P., AND J.R., S. Classification of hyperspectral data from urban areas based on extended morphological profiles. *Geoscience and Remote Sensing, IEEE Transactions on* 43, 3 (march 2005), 480 – 491.

- [11] BIANCONI, F., AND FERNÁNDEZ, A. Evaluation of the effects of gabor filter parameters on texture classification. *Pattern Recognition* 40 (2007), 3325–3335.
- [12] BURGESS, C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 2 (1998), 121–167.
- [13] CAMPS-VALLS, G., AND BRUZZONE, L. Kernel-based methods for hyperspectral image classification. *IEEE Trans. on Geoscience and Remote Sensing* 43 (2005), 1351–1362.
- [14] CHANGQING, Z., AND XIAOMEI, Y. Study of remote sensing image texture analysis and classification using wavelet. *International Journal of Remote Sensing* 19, 16 (1998), 3197–3203.
- [15] CHEN, C., AND HO, P. Statistical pattern recognition in remote sensing. *Pattern Recognition* 41 (2008), 2731–2741.
- [16] CHEN, C.-C., AND CHEN, C.-C. Filtering methods for texture discrimination. *Pattern Recognition Letters* 20, 8 (1999), 783 – 790.
- [17] CHENG, Q., VARSHNEY, P., AND ARORA, M. Logistic regression for feature selection and soft classification of remote sensing data. *Geoscience and Remote Sensing Letters, IEEE* 3, 4 (2006), 491–494.
- [18] CHENG, Y. Mean shift, mode seek, and clustering. *IEEE Trans. on PAM* (1995).
- [19] CHI, M., AND BRUZZONE, L. Semisupervised classification of hyperspectral images by svms optimized in the primal. *Geoscience and Remote Sensing, IEEE Transactions on* 45, 6 (2007), 1870–1880.
- [20] CHI, M., AND YU, X. H. S. Mixture model label propagation. In *19th ACM international conference on Information and knowledge management* (October 2010), pp. 1889 – 1892.
- [21] COCKS, T., JENSSEN, R., STEWART, A., WILSON, I., AND SHIELDS, T. The hymap airborne hyperspectral sensor: The system, calibration and performance. In *Proc. of First EARSEL Workshop on Imaging Spectroscopy* (1998).
- [22] COMANICIU, D., AND MEER, P. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 5 (may 2002), 603 –619.
- [23] D.A., L. Multispectral land sensing: where from, where to? *IEEE Transactions on Geoscience and Remote Sensing* 43, 3 (2005), 414–421.
- [24] DAUBECHIES, I. The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on* 36, 5 (1990), 961–1005.
- [25] DEMIR, B., AND BRUZZONE, L. A novel active learning method for support vector regression to estimate biophysical parameters from remotely sensed images. In *SPIE Remote Sensing* (2012), pp. 85370L–85370L–8.



- [26] DOPIDO, I., LI, J., PLAZA, A., AND BIOUCAS-DIAS, J. Semi-supervised active learning for urban hyperspectral image classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International* (2012), pp. 1586–1589.
- [27] DUDA, R., AND HART, P. *Pattern classification*. John-Wiley and Sons, 2001.
- [28] DUIN, R., DE RIDDER, D., JUSZCZAK, P., LAI, C., PACLIK, P., PEKALSKA, E., AND TAX, D. Prtools4, 2010.
- [29] DUIN, R., FRED, A., LOOG, M., AND PEKALSKA, E. Mode seeking clustering by knn and mean shift evaluated. In *SSPR SPR 2012 Lecture Notes in Computer Science* (2012), vol. 7626, Springer, pp. 51–59.
- [30] ET AL., A. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment* 113 (2009), 110–122.
- [31] ET AL., B. S. The civil air patrol archer hyperspectral sensor system. In *Proc. of Airborne ISR Systems and Applications II, SPIE* (2005), vol. 5787, pp. 17–28.
- [32] F., D., P., G., A., F., J.A., P., J.A., B., AND K., A. Exploiting spectral and spatial information in hyperspectral urban data with high resolution. *Geoscience and Remote Sensing Letters, IEEE* 1, 4 (oct. 2004), 322 – 326.
- [33] F., M., AND L., B. Classification of hyperspectral remotesensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing* 42, 8 (2004), 1778–1790.
- [34] F., T., C.-K., C., AND G.-R., L. Texture analysis for three dimension remote sensing data by 3d glcm. In *27th Asian Conference on Remote Sensing* (August 2006), pp. 1 – 6.
- [35] FAUVEL, M., CHANUSSOT, J., AND BENEDIKTSSON, J. Evaluation of kernels of kernels for multiclass classification of hyperspectral remote sensing data. In *Proceedings of ICASSP* (2006), pp. 813–816.
- [36] FERECATU, M., AND BOUEMAA, N. Interactive remote-sensing image retrieval using active relevance feedback. *Geoscience and Remote Sensing, IEEE Transactions on* 45, 4 (2007), 818–826.
- [37] FILIPPONE, M., CAMASTRA, F., MASULLI, F., AND ROVETTA, S. A survey of kernel and spectral methods for clustering. *Pattern Recognition* 41, 1 (2008), 176 – 190.
- [38] FLEISS, J. *Statistical methods for rates and proportions*. John-Wiley and Sons, 1981.
- [39] FOGEL, I., AND SAGI, D. Gabor filters as texture discriminator. *Biological Cybernetics* 61 (1989), 103–113.
- [40] FOODY, G. M. Status of land cover classification accuracy assessment. *Remote sensing of environment* 90 (2002), 185–201.

- [41] FUKUNAGA, K., AND HOSTETLER, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory* 21, 1 (1977), 32–40.
- [42] G., C.-V., AND L., B. Kernel-based methods for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on* 43, 6 (june 2005), 1351 – 1362.
- [43] G., C.-V., L., G.-C., J., M.-M., J., V.-F., AND J., C.-M. Composite kernels for hyperspectral image classification. *Geoscience and Remote Sensing Letters, IEEE* 3, 1 (jan. 2006), 93 – 97.
- [44] G., M., AND M., L. Support vector machines for hyperspectral image classification with spectral-based kernels. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International (july 2003)*, vol. 1, pp. 288 – 290 vol.1.
- [45] G., R., X., D., F., F., AND J., Z. Texture feature analysis using a gauss-markov model in hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on* 42, 7 (july 2004), 1543 – 1551.
- [46] GIACINTO, G., ROLI, F., AND VERNAZZA, G. *Comparison and combination of statistical and Neural Network algorithms for remote-sensing image classification*. Springer-Verlag, 1997, pp. 117–124.
- [47] GUALTIERI, J. A., AND CROMP, R. F. Support vector machines for hyperspectral remote sensing classification. In *Proc. SPIE* (1998), vol. 3584, pp. 221–232.
- [48] HARALICK, R., SHANMUGAM, K., AND DINSTEIN, I. Texture features for image classification. *IEEE Trans. Systems, Man, and Cybernetics* 3 (1973), 610–621.
- [49] HUANG, C., DAVIS, L. S., AND TOWNSHEND, J. R. G. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* 23, 4 (2002), 725–749.
- [50] HUGHES, G. F. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. on Information Theory* 14 (1968), 55–63.
- [51] J.A., B., PESARESI, M., AND AMASON, K. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *Geoscience and Remote Sensing, IEEE Transactions on* 41, 9 (sept. 2003), 1940 – 1949.
- [52] J.A., R., AND X., J. *Remote Sensing Digital Image Analysis: An Introduction*. Springer, 2006.
- [53] JACKSON, Q., AND D.A., L. Adaptive bayesian contextual classification based on markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on* 40, 11 (nov 2002), 2454 – 2463.

- [54] JAIN, A., AND HEALEY, G. A multiscale representation including opponent color features for texture recognition. *IEEE Trans. on Image Processing* 7 (1998), 124–128.
- [55] JAIN, A., DUIN, R., AND MAO, J. Statistical pattern recognition: a review. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22, 1 (jan 2000), 4–37.
- [56] JAIN, A., AND FARROKHNI, F. Unsupervised texture segmentation using gabor filters. *Pattern Recognition* 24 (1991), 1167–1186.
- [57] JIMENEZ, L., AND LANDGREBE, D. Hyperspectral data analysis and supervised feature reduction via projection pursuit. *IEEE Trans. on Geoscience and Remote Sensing* 37, 6 (Nov. 1999), 2653–2667.
- [58] JIMENEZ-RODRIGUEZ, L., ARZUAGA-CRUZ, E., AND VELEZ-REYES, M. Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data. *Geoscience and Remote Sensing, IEEE Transactions on* 45, 2 (2007), 469–483.
- [59] K., B. Multispec, 2012.
- [60] KOONTZ, W., NARENDRA, P., AND FUKUNAGA, K. A graph-theoretic approach to non-parametric cluster analysis. *IEEE Transactions on Computer* 25 (1976), 936–944.
- [61] KUMAR, S., GHOSH, J., AND CRAWFORD, M. Best-bases feature extraction algorithms for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on* 39, 7 (2001), 1368–1379.
- [62] L., B., CHI, M., AND M., M. A novel transductive svm for semisupervised classification of remote-sensing images. *Geoscience and Remote Sensing, IEEE Transactions on* 44, 11 (nov. 2006), 3363–3373.
- [63] LANDGREBE, D. A. *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken and NJ: Wiley, 2003.
- [64] LI, J. *Discriminative image segmentation: applications to hyperspectral data*. PhD thesis, Universidade Tecnica de Lisboa, 2011.
- [65] LI, J., BIOUCAS-DIAS, J., AND PLAZA, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *Geoscience and Remote Sensing, IEEE Transactions on* 48, 11 (nov. 2010), 4085–4098.
- [66] M., D. M., J.A., B., J., C., AND L., B. The evolution of the morphological profile: from panchromatic to hyperspectral images. In *Optical Remote Sensing*, vol. 3. Springer Berlin Heidelberg, 2011, pp. 123–146.
- [67] M., F. *Spectral and Spatial Methods for the Classification of Urban Remote Sensing Data*. PhD thesis, Grenoble Institute of Technology and University of Iceland, 2007.

- [68] M., F., J.A., B., J., C., AND J.R., S. Spectral and spatial classification of hyperspectral data using svms and morphological profiles. *Geoscience and Remote Sensing, IEEE Transactions on* 46, 11 (nov. 2008), 3804–3814.
- [69] M., P., AND J.A., B. A new approach for the morphological segmentation of high-resolution satellite imagery. *Geoscience and Remote Sensing, IEEE Transactions on* 39, 2 (feb 2001), 309–320.
- [70] MA, W., AND MANJUNATH, B. A comparison of wavelet transform features for texture image annotation. In *Image Processing, 1995. Proceedings., International Conference on* (1995), vol. 2, pp. 256–259 vol.2.
- [71] MALLAT, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 11 (1989), 674–693.
- [72] MARTÍNEZ-USÓ, A., PLA, F., AND GARCÍA-SEVILLA, P. Clustering-based hyperspectral band selection using information measures. *IEEE Trans. on Geoscience & Remote Sensing* 45 (2007), 4158–4171.
- [73] MELGANI, F., AND BRUZZONE, L. Classification of hyperspectral remote sensing images with support vector machines. *Geoscience and Remote Sensing, IEEE Transactions on* 42, 8 (2004), 1778–1790.
- [74] NG, A., JORDAN, M., AND WEISS, Y. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 14, 5 (2002), 849–856.
- [75] O., C., B., S., AND A., Z. *Semi-supervised learning*. MIT Press, 2010.
- [76] OJALA, T., PIETIKAINEN, M., AND MAAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI* 24 (2002), 971–987.
- [77] P., G. A collection of data for urban area characterization. In *Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004 IEEE International* (sept. 2004), vol. 1, pp. 7 vol. (cviii+4896).
- [78] PAL, N., AND PAL, S. A review on image segmentation techniques. *Pattern Recognition* 26 (1993), 1277 – 1294.
- [79] PEARLMAN, J., SEGAL, C., LIAO, L., CARMAN, S., FOLKMAN, M., BROWNE, B., ONG, L., AND UNGAR, S. Development and operations of the eo-1 hyperion imaging spectrometer. In *Proc. of Earth Observing Systems V SPIE* (2000), vol. 4135, pp. 243 – 253.
- [80] PERSELLO, C. Interactive domain adaptation for the classification of remote sensing images using active learning. *Geoscience and Remote Sensing Letters, IEEE* 10, 4 (2013), 736–740.

- 
- [81] PETROU, M., AND GARCÍA-SEVILLA, P. *Image Processing: Dealing with Texture*. John-Wiley and Sons, 2006.
- [82] PICHLER, O., TEUNER, A., AND HOSTICKA, B. J. A comparison of texture feature extraction using adaptive gabor filtering, pyramidal and tree structured wavelet transforms. *Pattern Recognition* 29, 5 (1996), 733 – 742.
- [83] PRESS, C. U. *Cambridge Advanced Learners Dictionary*. Cambridge University Press, 2008.
- [84] PRINCE, S. J. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.
- [85] RICKARD, L. J., BASEDOW, R. W., ZALEWSKI, E. F., SILVERGLATE, P. R., AND LANDERS, M. Hydice: an airborne system for hyperspectral imaging. In *Proc. of Imaging Spectrometry of the Terrestrial Environment, SPIE* (1993), vol. 1937, pp. 173–179.
- [86] R.O., G. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment* 65, 3 (1998), 227 – 248.
- [87] ROBERTS, S. J. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition* 30, 2 (1997), 261 – 272.
- [88] SÁNCHEZ, J. S., MOLLINEDA, R. A., AND SOTUCA, J. M. An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis and Applications* 10 (2007), 189 – 201.
- [89] SERPICO, S., AND BRUZZONE, L. A new search algorithm for feature selection in hyperspectral remote sensing images. *Geoscience and Remote Sensing, IEEE Transactions on* 39, 7 (2001), 1360–1367.
- [90] SERPICO, S., AND MOSER, G. Extraction of spectral channels from hyperspectral images for classification purposes. *Geoscience and Remote Sensing, IEEE Transactions on* 45, 2 (2007), 484–495.
- [91] SETTLES, B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- [92] SHAPIRO, L.G., AND STOCKMAN, G. *Computer Vision*. Prentice-Hall, 2001.
- [93] SHI, M., AND HEALEY, G. Hyperspectral texture recognition using a multiscale opponent representation. *IEEE Trans. on Geoscience and Remote Sensing* 41 (2003), 1090–1095.
- [94] SIEDLECKI, W., AND SKLANSKY, J. On automatic feature selection. *International Journal of Pattern Recognition and Artificial Intelligence* 02, 02 (1988), 197–220.

- [95] SWAIN, P., AND DAVIS, S. *Remote sensing: The quantitative approach*. Advanced book program. McGraw-Hill International Book Co., 1978.
- [96] T., C., AND P., H. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on* 13, 1 (january 1967), 21 –27.
- [97] TARABALKA, Y. *Classification of hyperspectral data using spectral-spatial approaches*. PhD thesis, University of Iceland and Institut Polytechnique de Grenoble, 2010.
- [98] TUIA, D., RATLE, F., PACIFICI, F., KANEVSKI, M., AND EMERY, W. Active learning methods for remote sensing image classification. *Geoscience and Remote Sensing, IEEE Transactions on* 47, 7 (july 2009), 2218 –2232.
- [99] TUIA, D., VOLPI, M., COPA, L., KANEVSKI, M., AND MUNOZ-MARI, J. A survey of active learning algorithms for supervised remote sensing image classification. *Selected Topics in Signal Processing, IEEE Journal of* 5, 3 (2011), 606–617.
- [100] V, V. *Remote sensing: models and methods for image processing*. Elsevier, 2007.
- [101] WANG, F., AND ZHANG, C. Label propagation through linear neighborhoods. *Knowledge and Data Engineering, IEEE Transactions on* 20, 1 (jan. 2008), 55 –67.
- [102] WANG, J., AND CHANG, C.-I. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *Geoscience and Remote Sensing, IEEE Transactions on* 44, 6 (2006), 1586–1600.
- [103] WILSON, R., AND SPANN, M. A new approach to clustering. *Pattern Recognition* 23, 12 (1990), 1413 – 1425.
- [104] XIN, H., AND LIANGPEI, Z. A comparative study of spatial approaches for urban mapping using hyperspectral rosis images over pavia city, northern italy. *International Journal of Remote Sensing* 30, 12 (2009), 3205–3221.
- [105] YANG, H., MEER, F., BAKKER, W., AND TAN, Z. A back-propagation neural network for mineralogical mapping from aviris data. *International Journal of Remote Sensing* 20 (1999), 97–110.
- [106] Y.TARABALKA, J.CHANUSSOT, AND J.A.BENEDIKTSSON. Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Trans. on Geoscience & Remote Sensing* 47, 8 (2009), 2973–2987.
- [107] Y.TARABALKA, J.CHANUSSOT, AND J.A.BENEDIKTSSON. Segmentation and classification of hyperspectral images using watershed transformation. *Patt.Recogn.* 43, 7 (2010), 2367–2379.

- 
- [108] Y.TARABALKA. J.CHANUSSOT, J. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Trans. Systems, Man, and Cybernetics* 40, 5 (2010), 1267–1279.
  - [109] ZHANG, X., N.H., Y., AND C.G., O. Wavelet domain statistical hyperspectral soil texture classification. *Geoscience and Remote Sensing, IEEE Transactions on* 43, 3 (march 2005), 615 – 618.
  - [110] ZHOU, H., MAO, Z., AND WANG, C. Classification of coastal areas by airborne hyperspectral image. In *Proceedings of SPIE* (2005), pp. 471–476.





# Appendix A

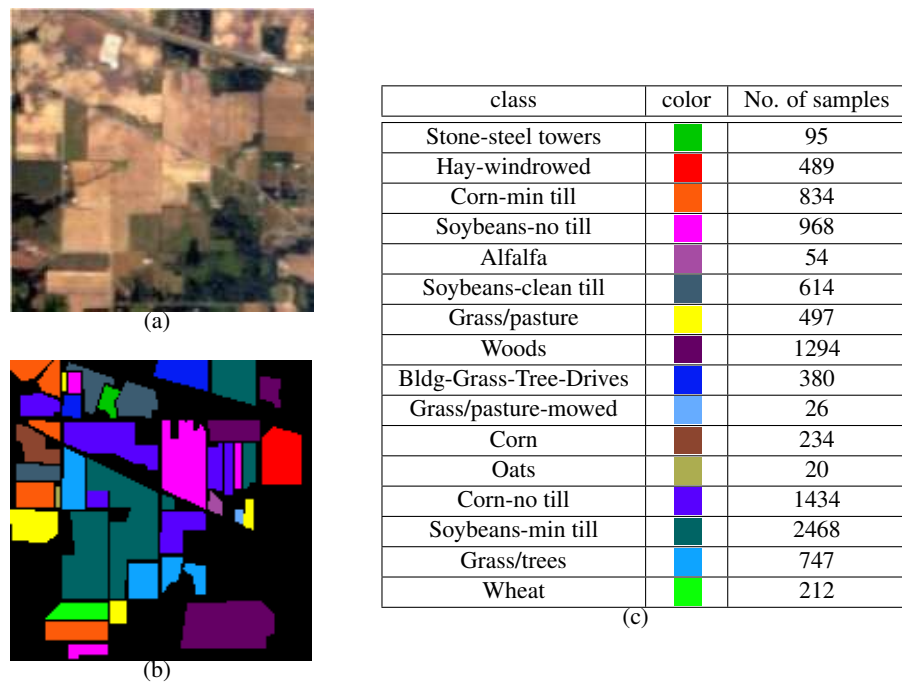
## Datasets

For the experimental phase, different hyper-spectral databases have been used in this thesis. For the three of them their groundtruth is available. This is an introduction of those data sets, their main characteristics and the classes they contain.

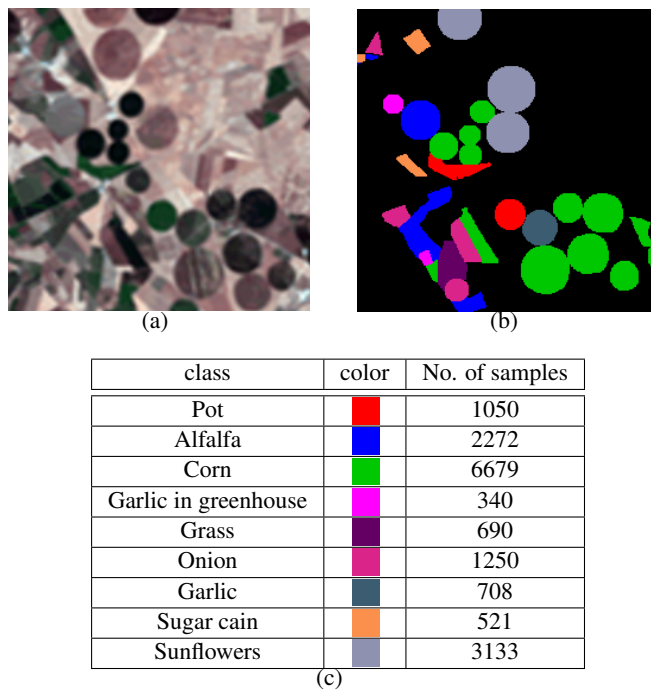
**AVIRIS** Hyper-spectral image 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. It is still available from [59]. The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. Figure A.1 shows the image together with the sixteen available classes that range from 20 to 2468 pixels in size. Due to the small size of some classes, one can find in literature that small classes are dismissed [30][43][57]. In this paper, those classes smaller than 400 pixels will be ignored in some experiments in order to compare our results with other authors. From the 220 bands that composed the image, 20 are usually ignored (the ones that cover the region of water absorption or with low SNR [57]).

**CHRIS-PROBA** This database is an acquisition of the PROBA satellite using CHRIS sensor, which has several operating modes. The image used here comes from the mode with a spatial resolution of 34m, obtaining a set of 62 spectral bands that range from 400 to 1050nm. It is  $641 \times 617$  pixels representing nine classes that are composed of crops and an unknown background class. A section of  $255 \times 255$  that contains all classes is shown in Figure A.2. Concretely, this image covers an area near to Barrax (Albacete, Spain). In this case, 52 bands remains when discarding the 10 lower SNR bands.

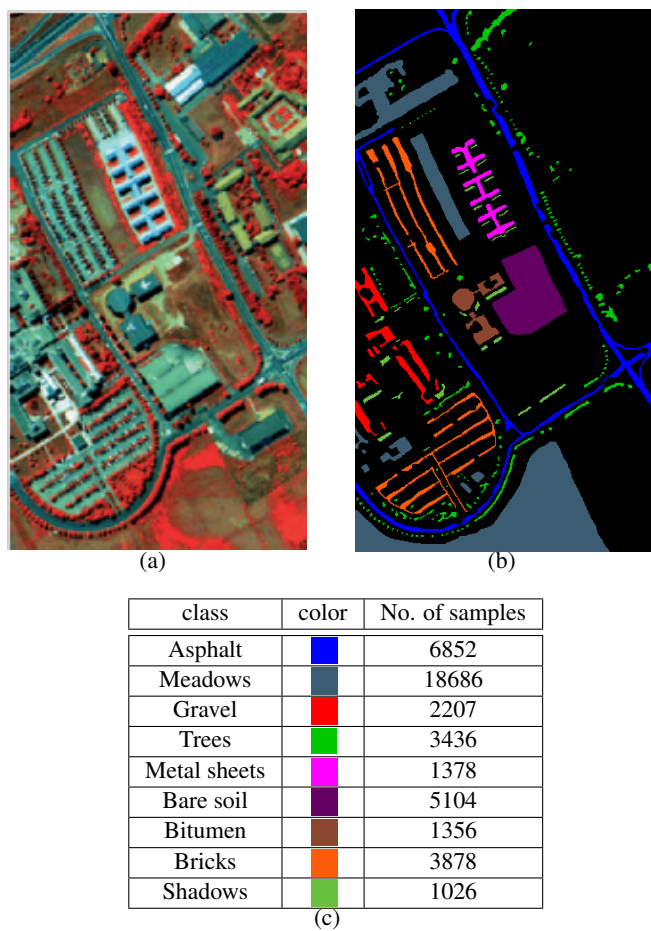
**ROSIS** The third dataset was collected in 2003 by the ROSIS sensor over the urban area of the University of Pavia, Italy. The image is  $610 \times 340$  pixels, with a spatial resolution of 1.3 m/pixel. The number of data channels in the acquired image is 115 (with a spectral range from 430 to 860 nm). The 12 most noisy channels are often removed, and the remaining 103 bands are used for experimentation. Nine ground-truth classes were considered in the experiments, which are shown in Figure A.3.



**Figure A.1:** Hyper-spectral image AVIRIS (92AV3C over the Indian Pines Test Site). (a) Color composition; (b) Ground-truth; (c) Target classes to be recognized.

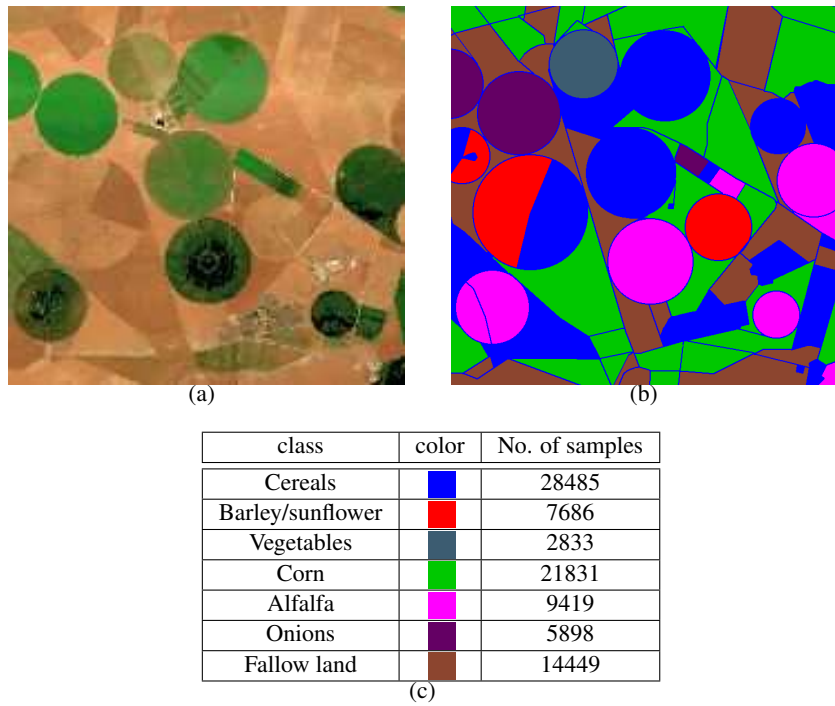


**Figure A.2:** Hyper-spectral image captured by the CHRIS-PROBA system. (a) Color composition; (b) Ground-truth; (c) Target classes.



**Figure A.3:** Hyper-spectral image captured by the ROSIS system at the University of Pavia. (a) Color composition; (b) Ground-truth; (c) Target classes.

**HYMAP** This dataset was collected within DAISEX'99 project (<http://www.uv.es/leo/daisex>). The HyMap Imaging Spectrometer is an Australian development built by Integrated Spectronics Ltd, and operated by HyVista Corp. The wavelength range between 400 and 2500 nm is covered in 126 spectral bands with a bandwidth of 16 nm. The scenario used here is an image of 301x301 over the area of Barrax (Albacete, Spain) and contains 7 classes as detailed in Figure A.4.



**Figure A.4:** Hyper-spectral image captured by the HyMap imaging spectrometer. (a) Color composition; (b) Ground-truth; (c) Target classes.

For preparing the data sets for experiments, correlation and size were decreased using a band selection method. Band selection was chosen over feature extraction because it preserve the original data. WaLuMI was chosen for preserving the original bands, providing as an output a subset of them. However, any other band selection method that fulfills similar criteria can be used instead.

In Table A.1 the set of bands selected for sets between 1 and 10 bands are stated. Note that the process is not incremental, this is because they are selected to have the highest amount of information and least correlation as a group. Varying the size of the set may lead to the fact the optimal set is not simply obtained by adding the next best band but finding another group that optimize that criteria.

no. of bands	selected bands	
	AVIRIS	CHRIS-PROBA
1	4	0
2	4, 67	0, 45
3	4, 67, 87	0, 20, 45
4	4, 67, 87, 128	0, 20, 45, 59
5	4, 67, 87, 129, 182	0, 20, 40, 46, 59
6	4, 51, 67, 87, 129, 182	0, 20, 30, 40, 46, 59
7	4, 51, 67, 78, 87, 129, 182	0, 9, 20, 30, 40, 46, 59
8	4, 51, 67, 78, 87, 99, 129, 182	0, 9, 20, 30, 40, 46, 56, 59
9	4, 24, 51, 67, 78, 87, 99, 129, 182	0, 9, 20, 30, 34, 40, 46, 56, 59
10	4, 24, 51, 67, 78, 87, 99, 118, 129, 182	0, 9, 17, 23, 30, 34, 40, 46, 56, 59

no. of bands	selected bands	
	ROSIS	HyMap
1	92	1
2	51, 92	31, 108
3	30, 53, 92	31, 108, 52
4	30, 53, 76, 92	1, 31, 52, 108
5	2, 30, 53, 76, 92	1, 31, 52, 79, 119
6	2, 20, 33, 53, 76, 92	1, 31, 52, 79, 91, 119
7	2, 20, 33, 45, 58, 76, 92	1, 28, 41, 52, 79, 107, 122
8	2, 20, 33, 45, 58, 71, 76, 92	1, 14, 28, 41, 52, 79, 107, 122
9	2, 11, 21, 33, 45, 58, 71, 76, 92	1, 14, 28, 41, 52, 69, 79, 107, 122
10	92, 58, 33, 76, 2, 21, 45, 71, 11, 1	1, 14, 28, 41, 52, 69, 79, 91, 107, 122

**Table A.1:** Selected bands using WaLuMi for AVIRIS, Chris-Proba, ROSIS and HyMap for  $B$  varying from 1 to 10. Note that the selection is not incremental but most of bands are often repeated.

# Appendix **B**

## Publications

## Textural Features for Hyperspectral Pixel Classification

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

Depto. Lenguajes y Sistemas Informáticos  
Jaume I University, Campus Riu Sec s/n 12071 Castellón, Spain  
{orajadel, pgarcia, pla}@lsi.uji.es  
<http://www.vision.uji.es>

**Abstract.** Hyperspectral remote sensing provides data in large amounts from a wide range of wavelengths in the spectrum and the possibility of distinguish subtle differences in the image. For this reason, the process of band selection to reduce redundant information is highly recommended to deal with them. Band selection methods pursue the reduction of the dimension of the data resulting in a subset of bands that preserves the most of information. The accuracy is given by the classification performance of the selected set of bands. Usually, pixel classification tasks using grey level values are used to validate the selection of bands. We prove that by using textural features, instead of grey level information, the number of hyperspectral bands can be significantly reduced and the accuracy for pixel classification tasks is improved. Several characterizations based on the frequency domain are presented which outperform grey level classification rates using a very small number of hyperspectral bands.

### 1 Introduction

Hyperspectral imagery consists of large amounts of channels covering the different wavelengths in the spectrum. These images represent a very rich source of information that allows an accurate recognition of the different areas to be obtained through the use of pattern classification techniques. For this reason, traditionally, this kind of images has been used in remote sensing applications. However, nowadays they are also widely used in medical imaging, product quality inspection or even fine arts. The main problems to deal with hyperspectral images are the high dimension of this data and its high correlation. In the context of supervised classification, an additional problem is the so-called Hughes phenomenon that occurs when the training set size is not large enough to ensure a reliable estimation of the classifier parameters. As a result, a significant reduction in the classification accuracy can be observed [3], [4], [5]. To overcome the Hughes phenomenon the original hyperspectral bands are considered as features and feature-reduction algorithms are applied [11]. They process the original set of features to generate a smaller size set of features with the aim of maximizing the classification accuracy. A particular class of feature reduction methods are band selection methods [9], [10], [7], which select a subset of the original set of bands and discard the remaining to reduce redundant information in the image representation without losing classification accuracy in a significant way. Methods of band selection obtain subsets of relevant bands so as to get the best classification performance. The performance of the



band selection is usually measured through pixel classification accuracy based on grey level pixel data.

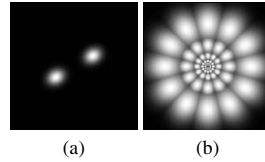
In this paper we propose the use of several frequential texture features to describe each individual pixel. The aim of this characterization is to reduce as much as possible the number of hyperspectral bands required in the global process while keeping the final pixel classification accuracy as high as possible. We start from the band selection scheme described in [7] and compare the classification accuracies obtained using grey level features against textural features. Gabor filters as well as wavelets features are considered in our study. Also, modified versions of Gabor filters are considered with two main objectives: obtaining a more detailed analysis of medium and high frequencies, and, simplifying their computational cost without decreasing their discriminant power.

## 2 Textural Features

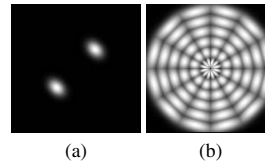
In hyperspectral imaging it is very common to characterize each pixel using a feature vector formed by the grey level values of that pixel in a given set of bands. To measure the performance of a band selection method, a series of pixels are characterized using their grey level values in the selected bands. The rate of correct classification obtained is compared to the classification rate obtained using the whole set of bands, to check the goodness of the selected group of bands as a representation of the entire hyperspectral image.

Now, our purpose is to describe the textural characteristics of a group of selected bands as they are supposed to portray the common features of pixels, that is, the texture they represent. For this reason, we have considered a series of frequential filters in order to extract features from the frequency domain to characterize pixels rewarding their textural features. In all cases, we consider a basic tessellation of the frequency domain taking into account several frequency bands and orientations [8]. A filter mask is applied over each area defined in the tessellation in order to select the frequencies within the chosen area. Then, for each area, we obtain its inverse Fourier transformation into the space domain. The result is an “image” which contains only frequencies in the chosen area, telling us which parts of the original image contain frequencies in this area. Repeating this process for all frequency areas we will have a stack of “images”. Therefore, for each pixel we have as many values as frequency areas we used, that is, one value per output “image”. This vector of values is used as the frequency signature of each pixel.

The first sort of filters considered are the well known Gabor filters. We construct a basic tessellation of the frequency domain considering several frequency bands and orientations. Each frequency band is double the previous one and a Gaussian mask is applied over each frequency area. Figure 1(a) shows an example of a Gabor filter considered. Figure 1(b) shows the maximum value of all Gabor filters considered for a given tessellation using four frequency bands and six different orientations. As it can be seen in this figure, each individual filter expands far away from the limit of the area defined in the tessellation. For this reason, two variations of these filters



**Fig. 1.** (a) An example of Gabor filter in the frequency domain (b) Maximum value of all Gabor filters considered for a given tessellation using four frequency bands and six different orientation

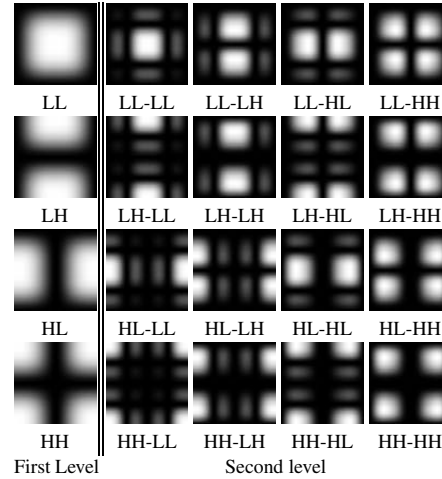


**Fig. 2.** (a) An example of filter keeping constant the width of the frequency band (b) Maximum value of all filters considered for a given tessellation using six frequency bands and six different orientation

(as described in [8]) could also be considered. First, similar Gaussians are applied over each frequency area, but truncating them beyond the limits of the areas in order to eliminate contributions of the same frequencies in different filters. On the other hand, the use of Gaussian masks over the frequency areas leads to the loss of importance of frequencies not lying nearby the center of these areas. That is why, also flat masks covering exactly each frequency area in the tessellation will be considered.

Another disadvantage in the application of Gabor filters using the basic tessellation scheme is that the frequency bands considered are not uniform. In this way, low frequencies are given more importance than middle or high frequencies. However, it is well known that texture information mainly falls in middle and high frequencies [1]. Therefore, we propose a detailed analysis of all frequencies by keeping constant the width of the frequency bands to be analyzed by each filter. Figure 2(a) shows an example of an individual filter using a complete Gaussian mask, while figure 2(b) shows the maximum value of all these filters considered for a given tessellation using six constant frequency bands and six different orientations. Note that, also in this case, truncated Gaussians and flat masks may be used.

Also features derived for each pixel using a wavelet decomposition will be considered. A wavelet decomposition is obtained using two appropriate filters: a low-pass filter  $L$  and a high-pass filter  $H$ . In this case, we have chosen to use a maximum overlap algorithm, that is, no subsampling is done. Therefore, after applying each filter, an image of the same size of the original image is obtained. Also, a wavelet packet analysis has been used, which means that not only low frequency components will be considered in further levels of analysis. In this case, all components will be taken into account. Figure 3 expresses the wavelet decomposition in the frequency domain for two levels of analysis using the Daubechies-4 filters.



**Fig. 3.** Wavelet decomposition expressed in the frequency domain for the two levels of analysis using the Daubechies-4 filters

### 3 Hyperspectral Database

The experimental results will consist of comparing the different characterization methods named above over a widely used hyperspectral database. The 92AV3C source of data corresponds to a spectral image, 145x145 pixel-sized, 220 bands, and 17 classes composed of different crop types, vegetation, man-made structures, and an unknown class. This image is acquired with the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data set and collected in June 1992 over the Indian Pine Test site in Northwestern Indiana [2].

### 4 Experimental Results

Experiments to characterize the texture of each pixel were run using all the textural features described before. For the basic tessellation (where each frequency band is double the other) four different bands and six different orientations (wedges of 30°) were considered, that is, a total of 24 features were used to characterize each pixel. For the constant tessellation, nine frequency bands of the same size and six directions were considered, which provide a total of 54 features for each pixel. These numbers of features are due to the symmetry of the Fourier transform when dealing with real numbers. For the wavelet decomposition, the Daubechies-4 filters were used until three levels of decomposition, providing a total of 84 features per pixel.

#### 4.1 Results Using the Best Band for Grey Level Classification

Let our band selection method be the one in [7], which has already proved its good performance for pixel classification using grey level features. It provides with a series of clusters, that is, sets of bands grouped depending on their mutual information. The bands that composed each set depend on the cluster number as every set by itself represents the best combination for all the possibilities. To test the discriminant power of each set of textural features we will run classification experiments using the only band in cluster number one.

The selection method reported band number 4 for the cluster of size one among the bands which compose the 92AV3C database. The results of all methods of characterization named above can be seen in Table 1. Classification has been performed using the K nearest neighbor rule with 3 neighbors. With classification purposes and due to the massive data that pixel characterization generates, samples have been divided into twenty independent sets keeping the a priori probability of each class and the k-nn rule has been used to classify all sets taken in pairs, one used as training set and the other as test set (1-1 knn3 method). Therefore, ten classification attempts have been performed without data dependencies among the attempts. In this way, a mean rate of all the attempts have been reported.

**Table 1.** Classification rates (in percentage) for the all characterization methods considered over band number 4 from 92AV3C database

Characterization method		Classification rate
Grey level values		18.85 %
Wavelet packets		27.77 %
Basic tessellation	Gauss	41.58 %
	Truncate	40.07 %
	Flat	41.31 %
Constant tessellation	Gauss	63.77 %
	Truncate	65.05 %
	Flat	65.78 %

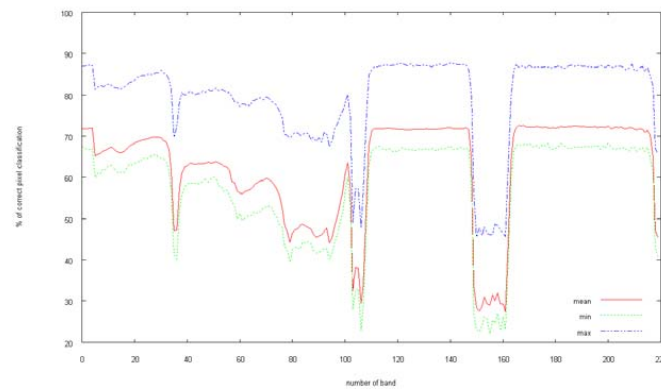
Results in table 1 show that all methods outperform grey level classification rates as it was obviously expected due to the higher number of features used. However, the wavelet features were worse than expected. It was the method that used the highest number of features and the percentage of correct classification was just a bit better than the grey level values. The basic tessellation performed significantly better than the wavelet features, using all sort of masks (Gaussian, truncate Gaussian, or flat). Finally, we can note that the constant tessellation outperformed the rest of features. When keeping the frequency band constant, the analysis is equally done for all frequencies bands what seems more appropriate for texture characterization. Moreover, we found that the sort of tessellation used influenced the final results much more than the sort of mask applied. Almost no difference was obtained when different masks were used. This is quite surprising as the used of truncate Gaussian masks should introduce important artifacts in the space domain, even more when the flat masks are considered. However,

the classification results are almost the same or even better when the flat masks were used. Perhaps, when Gaussian masks are used, frequencies do not equally contribute to the characterization and some of them lose their characterization significance. Thus, applying flat masks allows all frequencies to contribute equally and uniquely the characterization providing very good results and requiring less computational effort.

#### 4.2 Results Using Individual Bands

Previously, we have seen that using flat filters with a constant tessellation provided the best results of all the characterization methods studied. In consequence, we are now going to test these features for all the bands that make up the 92AV3C database.

Figure 4 shows the maximum, minimum, and mean percentage of correct classification for the same ten independent classification experiments described before run for each band in the database.



**Fig. 4.** Classification rates for each band of the 92AV3C database

From figure 4 we can observe that the maximum classification accuracy is not obtained at band number 4 as the chosen band selection method suggested. However, there are several bands, such as 171, that got better performance and consequently are more convenient. These results show that the textural features may be taken into account from the beginning in the band selection process, at least, as a testing criterion.

We can also notice in figure 4 that there are significant differences in the percentage of correct classification between bands. It is well known that several bands in the 92AV3C database are generally dismissed due to their low signal-to-noise ratio. These ranges are known to be bands 0 – 3, 102 – 109, 148 – 164, 216 – 219, as described in [6]. All these ranges provided the worst classification results, except for the range 0 – 3 which provided similar results to other bands. If these bands were not considered, even the worst band would provide quite good classification results taking into account that only one band is being used in each case.

### 4.3 Results for Clusters of Bands

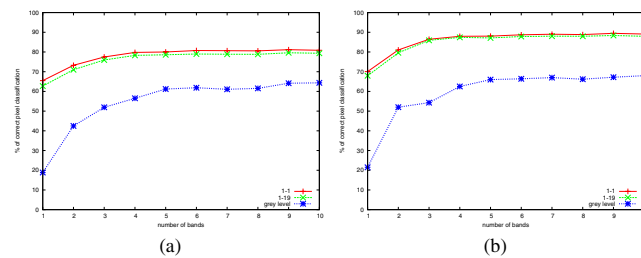
In this section we will show that textural characterization improves itself by using a higher number of bands even when the clusters of bands selected could not be optimal for these features, as it has been previously seen.

When more than one band is considered, all possible pairs of bands will be taken into account and textural features will be derived from them. Taking each pair of bands, a complex band will be formed using one of them as the real part and the other one as the complex part. When the Fourier transform is computed for these complex bands, the symmetric property is no longer fulfilled. Consequently, the number of filters to apply over each complex band doubles since each of the previous filter must be split into two parts due to the non-symmetrical transform.

The feature set obtained for each cluster will be divided into twenty random sets keeping the a priori probability of each class. In first place, as described in the previous sections, classification has been performed with the k-nn rule using 3 neighbors using pairs of sets, one used as training set and the other as test set (1-1 knn3 method). Other classification experiment consists of using each set once as training set whereas all the rest 19 sets are joined together to be used as a test set (1-19 knn3 method). In both experiments, a mean, maximum and minimum rate may be calculated, with ten and twenty independent attempts, respectively. For our current purpose, only the mean will be representative of our results and compared with classification rates reported in [7].

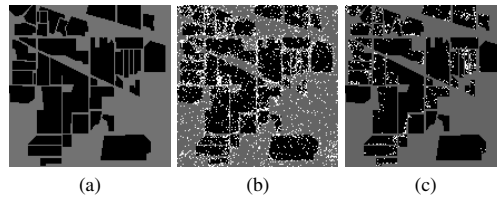
92AV3C database contains 17 different classes of textures, among them, the background class is composed by a heterogenous mixture of non-classified classes. Including this class into the classification process may confuse and decrease the performance rate due to its heterogenous nature, as different characterizations are assigned to the same class. For this reason, the more representative the characterization is of a class the less the classifier will fail, as pixels with a specific class will be properly characterized and so properly fit into its class by the classifier.

Fig. 5(a) shows classification results including the background class while Fig. 5(b) shows similar results without using the background class, in both cases for different numbers of bands in the cluster. It could be noted that textural features reaches



**Fig. 5.** Classification rates for clusters of 92AV3C database, with two methods of classification and compared with grey level characterization (a) taking into account the heterogenous class of background (b) without background class

stability sooner than the old method does, which means that a smaller number of bands is required in the whole process to reach a higher performance. As expected, when background is not taken into account performance enhances since background mistakes are removed (see Fig. 5(b)).



**Fig. 6.** (a) Ground truth of the 92AV3C database. Localization of classes in the space. (b) Maps of misclassified pixels (in white) using a clusters of 5 bands classified with a 1-1 knn3 method including the background class (c) Same map without the background class.

Fig. 6 presents the classification errors for the cluster composed by 5 bands. It shows misclassified pixels (white) by representing them in the space of the image superposed by the image's ground truth in order to distinguish the original classes recognized in the real image. Notice that the majority of the mistakes will be due to the heterogenous class or the proximity to it (see Fig. 6(b)). To avoid mistakes due to the background class and being able to analyze mistakes of the known classes, the background class may be ignored (observe Fig. 6(c)). In this case, misclassified pixels may be easily recognized and classification rates increase. Note that misclassified pixels fall mainly in the borders of the regions.

## 5 Conclusions

Results of hyperspectral texture characterization using several frequential filters has been presented in order to test band selection methods and reduce significantly the number of bands required in pixel classification tasks while improve the classification rates. Constant frequency band tessellation performed significantly better than traditional tessellation and the different masks tested performed similarly. We have chosen the flat masks due to its low computational cost. Different classification experiments have shown the stability of the textural features over different spectral bands, as well as when they were obtained from individual bands or from complex bands. Band selection methods usually take grey level pixel characterization as the validation criteria for their selection. We have shown that other validations should be taken into account as better classification rates may be obtained with textural information.

## Acknowledgments

This work has been partly supported by Fundació Caixa Castelló-Bancaixa through grant FPI PREDOC/2007/20 and project P1-1B2007-48, project CSD2007 00018 from Consolider Ingenio 2010, and project AYA2008-05965-C04-04 from Spanish CICYT.

## References

1. Chang, T., Kuo, C.C.J.: Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. Image Process.* 2, 429–441 (1993)
2. Freeware Multispectral Image Data Analysis System, <http://dynamo.ecn.purdue.edu/~biehl/MultiSpec>
3. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic, New York (1990)
4. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* 14(1), 55–63 (1968)
5. Jimenez, L.O., Landgrebe, D.A.: Supervised classification in highdimensional space: Geometrical, statistical, and symptotically properties of multivariate data. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.* 28(1), 39–54 (1998)
6. Landgrebe, D.A.: *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, Hoboken (2003)
7. Martínez-Usó, A., Pla, F., Sotoca, J.M., García-Sevilla, P.: Clustering-based Hyperspectral Band selection using Information Measures. *IEEE Transactions on Geoscience & Remote Sensing* 45(12), 4158–4171 (2007)
8. Petrou, M., García-Sevilla, P.: *Image Processing: Dealing with Texture*. John Wiley & Sons, Chichester (2006)
9. Richards, J., Jia, X.: *Remote Sensing Digital Image Analysis*, 3rd edn. Springer, Berlin (1999)
10. Serpico, S.B., Bruzzone, L.: A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 39(7), 1360–1367 (1994)
11. Shaw, G., Manolakis, D.: Signal processing for hyperspectral image explotation. *IEEE Signal Process. Mag.* 19(1), 12 (2002)



## On the Influence of Spatial Information for Hyper-spectral Satellite Imaging Characterization

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

Depto. Lenguajes y Sistemas Informáticos  
Jaume I University, Campus Riu Sec s/n 12071 Castellón, Spain  
{orajadel,pgarcia,pla}@lsi.uji.es

**Abstract.** Land-use classification for hyper-spectral satellite images requires a previous step of pixel characterization. In the easiest case, each pixel is characterized by its spectral curve. The improvement of the spectral and spatial resolution in hyper-spectral sensors has led to very large data sets. Some researches have focused on better classifiers that can handle big amounts of data. Others have faced the problem of band selection to reduce the dimensionality of the feature space. However, thanks to the improvement in the spatial resolution of the sensors, spatial information may also provide new features for hyper-spectral satellite data. Here, an study on the influence of spectral-spatial features combined with an unsupervised band selection method is presented. The results show that it is possible to reduce very significantly the number of spectral bands required while having an adequate description of the spectral-spatial characteristics of the image for pixel classification tasks.

### 1 Introduction

Hyper-spectral images are the results of a detailed measurement of the spectra acquired by a special sensor. Currently, some sensors can easily cover a spectral resolution of 10nm with a considerably high spatial resolution that can reach 1m per pixel for satellite images. As a result, hyper-spectral images are composed by a very high number of correlated bands (between 200 and 500 spectral bands). Dealing with this type of images means facing a very high dimensional problem.

Since the usage of the whole hyper-spectral data set can fall into the course of dimensionality, several band selection methods have been studied in order to avoid the large amount of correlated data, while keeping the discrimination between land cover classes [1].

When the spatial resolution in hyper-spectral images was not high enough, major efforts to improve pixel classification were done focusing at the classification stage by simply using the spectral features provided by the sensors. These type of processing often used neural networks [2], decision trees [3], Bayesian estimation [4] and kernel-based methods [5] for the classification of the pixels in the images. In particular, Support Vector Machines proved to obtain good performances in this task [6].

Because of the increase in the spatial resolution, spectral-spatial analysis has been lately an issue of high interest for the improvement of hyper-spectral imaging characterization [7] which is widely used for tasks like land-cover classification and segmentation of remote sensing images. Some basic spatial features have been used like the

mean value of a  $N \times N$  window around a pixel, the standard deviation of the values in this window, and the combination of the mean and standard deviation for a series of window sizes [6]. On the other hand, textural analysis has been widely discussed to study the spatial relationships in an image. This sort of features could be applied over hyper-spectral images in order to have a better description of the spectral-spatial properties. There exists a huge variety of methods [8]: co-occurrence matrices, wavelet analysis, Gabor filtering, Local Binary Patterns, etc.

It is likely that improving the characterization of the image may help to reduce even more the amount of spectral bands required for the classification task. To pursue this goal, we have chosen two different spectral-spatial characterization methods. In first place, simple statistics (mean and standard deviation) of the  $N \times N$  neighbors around a pixel will be considered for each spectral band. Later, a Gabor filter bank will be used to obtain features to describe the pixel in each band. Spectral-spatial feature extraction will be presented in Section 2. The hyper-spectral database used in our experiments is described in Section 3. The spectral-spatial methods proposed provide an improvement over the spectral classification as will be shown in Section 4. Finally, we draw out conclusions in Section 5.

## 2 Integration of Spatial Information in Imaging Characterization Methods

Pixel characterization aims at obtaining one feature vector for each pixel to be used in a pixel classification task in a multidimensional space. When only spectral data is used, the feature vector for every pixel is defined as the spectral curve provided by the sensor. The target of a spectral-spatial characterization method is to calculate a feature vector using the spectral data given and this can whether replace the spectral feature vector or being combined with it.

Let  $I^i(x, y)$  be the  $i^{th}$  band of an image containing  $B$  bands. When the spectral curve is used as the feature vector for each pixel in the image this vector is simply composed of the values provided by the sensor, that is:

$$\psi_{x,y} = \left\{ I^i(x, y) \right\}_{i=1}^B \quad (1)$$

### 2.1 Basic Spatial Characterization

Spectral-spatial analysis of the image is based on a series of values extracted from spatial operations involving its neighbor pixels (spatial features) [9]. Frequently the two statistics used are the mean and the standard deviation of the neighborhood. This is a very simple method to include spatial information obtaining only 2 features per pixel [6].

Let  $M_n^i(x, y)$  be the window  $n \times n$  centered in the pixel  $(x, y)$  of the spectral band  $i$ . Then, the feature vector for this pixel is defined by:

$$\phi_{x,y} = \left\{ mean(M_n^i(x, y)), standard\_deviation(M_n^i(x, y)) \right\}_{i=1}^B \quad (2)$$

462 O. Rajadell, P. García-Sevilla, and F. Pla

It is also possible to concatenate the features calculated from several window sizes (i.e.  $n = 3, 5, 7, 9$ ) increasing the size of the vector  $\phi$  depending on the number of windows used. This provides a multi-scale or multi-resolution description of the image.

## 2.2 Feature Extraction Based on Gabor Filters

Several features have been suggested in the literature for the description of texture information [8]. In this paper Gabor filtering will be used because, in general, they have provided the best results in different sort of texture characterization experiments [10] [11]. In this case, features are obtained by filtering the input image with a set of filters. The set of outputs obtained for each pixel in the image forms its feature vector. Here, the filter bank is defined to be a set of two-dimensional Gabor filters. Each Gabor filter is characterized by a preferred orientation and a preferred spatial frequency (scale). The filter acts as a local band-pass filter with optimal joint localization properties in the spatial domain and the frequency domain [12].

Gabor filters consist essentially of sine and cosine functions modulated by a Gaussian envelope. They can be defined by equation (3) where  $m$  is the index for the scale,  $n$  for the orientation and  $u_m$  is the central frequency of the scale [12].

$$f_{mn}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_m x \cos \theta_n + u_m y \sin \theta_n)) \quad (3)$$

Notice that set the condition  $f_{mn}(0, 0) = 0$  dismisses completely the effect of the measurements themselves and making the analysis independent from the pixel spectral values themselves.

Note that Gabor filters will be used in this case as a multi-dimensional extension of the technique designed for mono-channel images. In this way, multi-spectral images will be simply decomposed into separated channels and the same feature extraction process will be performed over each channel as shows equation (4).

$$h_{mn}^i(x, y) = I^i(x, y) * f_{mn}(x, y) \quad (4)$$

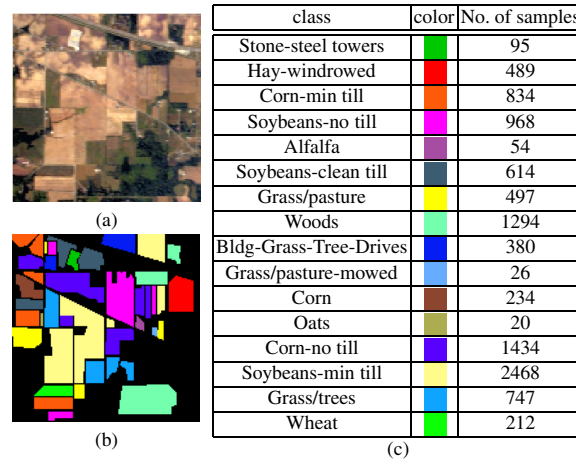
These responses are used to construct the final feature vector for each pixel.

$$\mathcal{Y}_{x,y} = \{h_{mn}^i(x, y)\}_{\forall i,m,n} \quad (5)$$

## 3 Hyper-spectral Data Set

To pursue the experimental campaign a widely used hyper-spectral database has been used, 92AV3C, known as AVIRIS. Figure 1 show a color composition, its corresponding ground-truth and the classes in it.

Hyper-spectral image data 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in North-western Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR [7]). The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. Fig. 1 shows the sixteen available classes, ranging from 20 to 2468 pixels in size. In it, three different growing states of soya can be found, together with other



**Fig. 1.** Hyper-spectral image AVIRIS (92AV3C). a)Color composition. b)Ground-truth. c)Target classes contained.

three different growing states of corn, woods, pasture and trees are the larger classes in terms of number of samples (pixels). Due to the small size of the rest of classes they are frequently dismissed in literature. In this paper, we will perform experiments using both 16 and 9 classes.

## 4 Spectral/Spatial Classification Results

As it has been pointed out, remote sensing has to deal with high dimensional feature vector where features are highly correlated. Consequently, band selection methods are frequently used. In our case, a band selection method presented by Martinez et al. in [1] has been used. Let  $D$  be a number of spectral bands such as  $D \leq B$ , where  $B$  is the total number of bands included in the database. This method provides the best set of  $D$  bands in term of uncorrelated information. It is based on a clustering approach that joins groups of bands depending on their mutual information. Once a partition of  $D$  groups is available, a representative band from each group is selected.

### 4.1 Classification Task

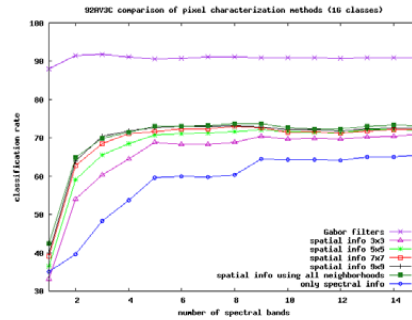
In Figures 2 and 3 a global view of the classification results using different spectral-spatial characterization methods can be found. The classification rates using only spectral information has also been included to be considered as a baseline reference. These results show the overall accuracy for four different sizes of windows to extract spatial information of the pixels ( $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ ,  $9 \times 9$ ), the combinations of these spatial features which is just a concatenation of all of them, and the Gabor textural features

464 O. Rajadell, P. García-Sevilla, and F. Pla

using two scales and four orientations. Every characterization method has been tested with the corresponding set of bands provided by the band selection algorithm from 1 to 15. Also the task with all bands in the dataset has been performed and can be observed in Table 1.

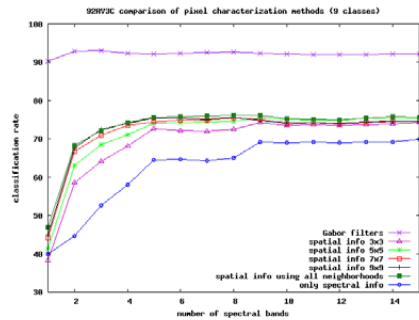
In these experiments, the pixels that form the whole image were divided into twenty non overlapping sets, keeping the a priori probability of each class. Therefore, no redundancies are introduced and each set is a representative set of the original image. The same sets of pixels are used in all experiments. Ten classification attempts were carried out with the k-nearest neighbor classifier with  $k = 3$  and the mean of the error rates of these attempts was taken as the final performance of the classifier for this experiment. Each classification attempt uses one of these sets for training and another set for testing. Each set is never used twice, so the attempts are totally independent.

Experiments using all 16 available classes are shown in Figure 2. As an alternative, experiments excluding the classes with a reduced number of samples have also been carried out using the same criterion as in [6]. Their results are presented in Figure 3. Better results, as expected, were got in this case. Small classes represent small structures in the image that are hard to recognize since their size is not enough to be capture by spatial features. Furthermore, some neighborhoods may be too big that several spatial structures could be considered at a time.



**Fig. 2.** Pixel classification rates for the 92AV3C database using all 16 classes. The number of spectral bands selected varies from 1 to 15.

Significant differences were obtained between spectral-spatial features and only spectral features even if the basic spatial features were used. Regarding these last sort of features, observe also that the larger the neighborhood used, the better classification results were obtained. Also, the concatenation of features obtained using different window sizes did not improve the results provided by using only the largest window. This means that, in this case, the spatial characterization is more reliable when we describe pixels by a fairly stable neighborhood. Furthermore, Gabor textural features outperformed all other methods very significantly. This points out that detailed spatial information is really discriminative for land use classification in this sort of images.



**Fig. 3.** Pixel classification rates for 92AV3C database using only the main 9 classes. The number of spectral bands selected varies from 1 to 15.

The differences between the characterization methods are not only due to the final classification rates obtained. Note also, that the number of spectral bands required to reach the stable behavior (where more spectral bands do not improve the classification results) is quite different. While spectral features require more than 12 bands, basic spatial features reach the stable zone with only 6–8 bands, while Gabor textural features required only 2–3 spectral bands.

In Table 1 the results obtained for several numbers of spectral bands can be compared with those obtain when using all 200 bands. Notice that, no matter the set of features used, no improvement is obtained by increasing the incoming data although the size of the problem is considerably increased.

**Table 1.** Accuracy for the 16 classes classification experiments of the 92AV3C dataset using different features. Results from the first sets of bands have been included together with the results obtained using the complete database (200 spectral bands).

# of spectral bands	Characterization methods			
	Spectral information	Spatial window 9 × 9	Spatial All windows	Gabor features
1	34.964	39.916	42.367	88.049
3	48.361	70.451	69.851	91.885
5	59.652	72.612	72.939	90.553
7	59.765	72.957	73.212	91.036
9	64.534	72.879	73.635	90.977
200	52.849	73.521	73.633	90.456

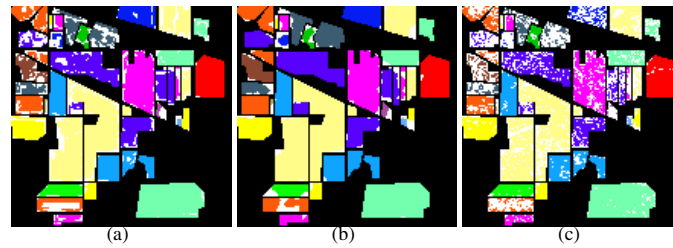
4.2 Segmentation Results

Since the problem we are tackling involve land-use pixel classification, the percentages of correct may not be enough to appreciate the goodness of the results. Pixel classification experiments assign a class label to each pixel in the test set. If we represent these

466 O. Rajadell, P. García-Sevilla, and F. Pla

labels in the position of their corresponding pixels we will obtain a segmentation map of the processed image. In Figure 4 this representation of the results is shown where misclassified pixels (errors) are represented in white color, while the rest of pixels are represented by their own class color as presented in the ground-truth shown in Figure 1. The results shown correspond to classification experiments where only one set of pixels was used for training (5% of the pixels in the image) and the other 19 sets of pixels were used for testing, using 10 spectral bands. Only the results for three characterization methods are shown. On the left, the results using the basic spatial features extracted from all different window sizes ( $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ ,  $9 \times 9$ ) are shown. Results using Gabor features are shown in the center of the figure. The results obtained using only spectral information are presented on the right.

Observe how the errors are distributed over the different classes. Spectral features (on the right) suffer from salt and pepper classification noise since the error is all over the areas and is not localized. However, when using Gabor textural features, the errors are located mainly in small areas and at the borders of the classes where the spatial features are mixing information from the heterogeneous background. We could say that the areas recognized using these features are more homogeneous. In the case of the basic spatial features, the errors are distributed in a similar way to the ones obtained using Gabor features but the results are worse in this case, so the misclassified pixels extend deeper inside the classes.



**Fig. 4.** Pixel classification results using 5% of the pixels for training for the 16 classes of the 92AV3C database, using 10 spectral bands. (a) Basic spatial features for all window sizes considered (b) Gabor textural features (c) Spectral features.

## 5 Conclusions

An experimental campaign over the 92AV3C dataset has been performed using several spectral-spatial characterization methods. Among them, the basic spatial features using simple statistics derived from a neighborhood and a Gabor textural features for a filter bank with two scales and four orientations have been used. Both basic and Gabor features outperform the naive spectral classification pointing out that taking advantage of the spatial resolution in the image is highly recommended for pixel classification tasks. Besides, Gabor textural features have provided very good classification results using a basic K-nearest neighbor classifier. Spectral features never provided results close to the

ones obtained using spatial information even when all two hundred spectral features were considered. In the segmentation experiments, spatial features have also proven their good performance providing quite homogenous regions and keeping the classification errors near the boundaries of the classes due to the influence of the heterogenous background. Furthermore, the good classification results obtained using spatial features required a minor number of spectral bands. Therefore, the use of spatial information can reduce the number of spectral bands required for pixel classification tasks and, at the same time, improve the rates of pixel classification.

### Acknowledgments

This work has been partly supported by grant FPI PREDOC/2007/20 from Fundació Caixa Castelló-Bancaixa and projects CSD2007-00018 (Consolider Ingenio 2010) and AYA2008-05965-C04-04 from the Spanish Ministry of Science and Innovation.

### References

1. Martínez-Usó, A., Pla, F., García-Sevilla, P.: Clustering-based hyperspectral band selection using information measures. *IEEE Trans. on Geoscience & Remote Sensing* 45, 4158–4171 (2007)
2. Yang, H., Meer, F., Bakker, W., Tan, Z.: A back-propagation neural network for mineralogical mapping from aviris data. *International Journal of Remote Sensing* 20, 97–110 (1999)
3. Zhou, H., Mao, Z., Wang, C.: Classification of coastal areas by airborne hyperspectral image. In: *Proceedings of SPIE*, pp. 471–476 (2005)
4. Chen, C., Ho, P.: Statistical pattern recognition in remote sensing. *Pattern Recognition* 41, 2731–2741 (2008)
5. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. *IEEE Trans. on Geoscience & Remote Sensing* 43, 1351–1362 (2005)
6. Plaza, A., et al.: Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment* 113, 110–122 (2009)
7. Landgrebe, D.A.: *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, Hoboken (2003)
8. Petrou, M., García-Sevilla, P.: *Image Processing: Dealing with Texture*. John-Wiley and Sons, West Sussex (2006)
9. Jimenez, L., Landgrebe, D.: Hyperspectral data analysis and supervised feature reduction via projection pursuit. *IEEE Trans. on Geoscience and Remote Sensing* 37(6), 2653–2667 (1999)
10. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18(8), 837–842 (1996)
11. Rajadell, O., García-Sevilla, P., Pla, F.: Filter banks for hyperspectral pixel classification of satellite images. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) *CIARP 2009*. LNCS, vol. 5856, pp. 1039–1046. Springer, Heidelberg (2009)
12. Fogel, I., Sagi, D.: Gabor filters as texture discriminator. *Biological Cybernetics* 61, 103–113 (1989)



## Scale Analysis of Several Filter Banks for Color Texture Classification

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

Depto. Lenguajes y Sistemas Informáticos  
Universitat Jaume I, Campus Riu Sec s/n 12071 Castellón, Spain  
{orajadel,pgarcia,pla}@lsi.uji.es  
<http://www.vision.uji.es>

**Abstract.** We present a study of the contribution of the different scales used by several feature extraction methods based on filter banks for color texture classification. Filter banks used for textural characterization purposes are usually designed using different scales and orientations in order to cover all the frequential domain. In this paper, two feature extraction methods are taken into account: Gabor filters over complex planes and color opponent features. Both techniques consider simultaneously the spatial and inter-channel interactions in order to improve the characterization based on individual channel analysis. The experimental results obtained show that Gabor filters over complex planes provide similar results to the ones obtained using color opponent features but using a reduced number of features. On the other hand, the scale analysis shows that some scales could be ignored in the feature extraction process without distorting the characterization obtained.

### 1 Introduction

Texture analysis has been tackled from different points of view in the literature. Literature survey provides us with a wide variety of well known texture analysis methods (co-occurrence matrices [5], wavelets [7], Gabor filters [3], local binary patterns [8], etc.) which have been mainly developed for grey level images.

Although the supremacy of filter-bank based methods for texture analysis have been challenged by several authors [12] [8] they are still one of the most frequently used methods for texture characterization. One goal of this paper is to analyze the influence of the scale parameter in several filter banks for texture analysis and study the information provided by each filter in order to reduce the characterization data required. Reducing the number of features used may make the feature extraction process easier.

It is well known that, when dealing with microtextures, the most discriminant information falls in medium and high frequencies [2] [9]. Therefore, it may be convenient to consider the influence of each frequency band separately in order to identify where the textural information could be localized.

Color texture analysis in multi-channel images has been generally faced as a multi-dimensional extension of techniques designed for mono-channel images. In this way, color images are decomposed into three separated channels and the same feature extraction process is performed over each channel. This definitely fails capturing the inter-channel properties of a multi-channel image.

On the other hand, in order to study these inter-channel interactions, color opponent features were proposed [6] which combine spatial information across spectral bands at different scales. Furthermore, we propose the use of similar features obtained using Gabor filters over complex planes which also try to describe the inter-channel properties of color textures, but using a smaller number of features.

The paper is organized as follows: first, the use of Gabor filters over complex channels and color opponent features are described in section 2. Section 3 describes the experiments performed and section 4 comments on the experimental results obtained. The conclusions are shown in section 5.

## 2 Feature Extraction

Let  $I^i(x, y)$  be the  $i^{th}$  channel of an image and  $f(x, y)$  a filter in the filter bank. The response of an image channel to the filter applied is given by:

$$h^i(x, y) = I^i(x, y) * f(x, y) \quad (1)$$

The response of a filter over an image channel may be represented by its total energy:

$$\mu_i = \sum_{x, y} h_i^2(x, y) \quad (2)$$

If a filter bank is applied, an image can be characterized by means of all the responses generated by all filters. It is possible to apply a filter in the space domain by a convolution or in the frequency domain by a product. In both cases, the feature obtained is the corresponding energy of the chosen group of pixels which responds to the filter applied [4].

When using filter banks, they are generally designed considering a dyadic tessellation of the frequency domain, that is, each frequency band (scale) considered is double the size of the previous one. It should not be ignored that this tessellation of the frequency domain thoroughly analyzes low frequencies, given less importance to medium and higher frequencies. Because the purpose of this work is to localize the texture information for color microtexture classification tasks, an alternative constant tessellation (giving the same width to all frequency bands) is proposed in order to ensure an equal analysis of all frequencies [10].

### 2.1 Gabor Filter Bank

Gabor filters consist essentially of sine and cosine functions modulated by a Gaussian envelope that achieve optimal joint localization in space and frequency [3]. They can be defined by eq. (3) and (4) where  $m$  is the index for the scale,  $n$  for the orientation and  $u_m$  is the central frequency of the scale.

$$f_{mn}^{real}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_m x \cos \theta_n + u_m y \sin \theta_n)) \quad (3)$$

$$f_{mn}^{imag}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \sin(2\pi(u_m x \cos \theta_n + u_m y \sin \theta_n)) \quad (4)$$

If symmetrical filters in the frequency domain are considered, only the real part of the filters in the space domain must be taken into account for convolution.

## 2.2 Complex Bands

If we filter each individual image channel we will lose all inter-channel information in the image. Hence, in order to take advantage of the inter-channel data, complex bands will be used instead. In this way, two real image channels are merged into one complex channel, one as the real part and the other one as the imaginary part. In this way we involve inter-channel information in each characterization process (similarly as the opponent features do, see next section). Since complex channels are no longer real, their corresponding FT is neither symmetrical. In this case, we suggest the usage of complex filters (non-symmetrical filters in the frequency domain).

As a result, for a cluster of image channels, we will consider all possible complex channels (pairs of channels). The Gabor filter bank will be applied over all complex channels as shown in eq. (5), where  $I^i(x, y)$  is the  $i^{th}$  image channel and  $f_{m,n}(x, y)$  the filter corresponding to the scale  $m$  and the orientation  $n$  in the filter bank previously defined.

$$h_{mn}^{ij}(x, y) = (I^i(x, y) + I^j(x, y)i) * f_{mn}(x, y) \quad (5)$$

The feature vector for each filter applied over the image is composed of the energy response to all filters in the filter bank, that is:

$$\psi_{x,y} = \{\mu_{mn}^{ij}(x, y)\}_{\forall i,j/i \neq j, \forall m,n} \quad (6)$$

As we are working with color images, the number of bands that compose the image is fixed to three. Even though, the size of the feature vector varies with the number of orientations and scales. For each complex channel, one feature is obtained for each filter applied what means that there will be as many features as filters for each complex band. So the total number of features is given by eq. (7) where  $M$  stands for the number of scales and  $N$  for the number of orientations.

$$size(\psi_{x,y}) = M \times N \times 3 \quad (7)$$

As inter-channel information is introduced in complex channels, it would be interesting to use some sort of decorrelation method (e.g. PCA) to minimize the correlation of RGB data in order to guarantee that merged information do introduce relevant information. Therefore, in the experiments carried out we will show results applying the filter bank directly over the RGB channels, and also over the PCA-RGB channels.

## 2.3 Opponent Features

Opponent features combine spatial information across image channels at different scales and are related to processes in human vision [6]. They are obtained from Gabor filters, computing firstly the difference of the outputs of two different filters. These differences among filters are needed for all pairs of image channels  $i, j$  with  $i \neq j$  and for all scales such that  $|m - m'| \leq 1$ :

$$d_{mm'n}^{ij}(x, y) = h_{mn}^i(x, y) - h_{m'n}^j(x, y) \quad (8)$$

512 O. Rajadell, P. García-Sevilla, and F. Pla

Then, the opponent features can be obtained as the energies of the computed differences:

$$\rho_{mm'n}^{ij} = \sum_{x,y} (d_{mm'n}^{ij}(x,y))^2 \quad (9)$$

In this way, opponent features use inter-channel information and minimize the correlation of channels which are expected to be highly correlated.

The feature vector for an image is the set of all opponent features for all image channels:

$$\varphi_{x,y} = \{\rho_{mm'n}^{ij}(x,y)\}_{\forall i,j/i \neq j, \forall m,m'/|m-m'| \leq 1, \forall n} \quad (10)$$

Hence, the size of the opponent feature vector also depends on the number of scales  $M$ , and orientations  $N/2$ , being the number of bands  $B = 3$  for color images. Note that the number of orientations used in this case is half the number of orientations used before. This is because now each filter is applied over a single image channel (that is, a real image) and, therefore, the other half filters will provide symmetrical responses.

$$size(\varphi_{x,y}) = \frac{size(\psi_{x,y})}{2} + B \times (B-1) \times (M-1) \times \frac{N}{2} \quad (11)$$

Note that, for usual values of  $M$  and  $N$ , the number of features is considerably increased in this case.

### 3 Experimental Setup

Several experiments have been conducted on texture classification in order to investigate the characterization properties of the filter banks described in previous sections. Also the effects of the different scales used to create the filter banks will be studied. Seventeen different color textures have been taken from the VisTex database [13] which are shown in Fig. 1. All of them are  $512 \times 512$  sized images that have been divided into sixty-four non-overlapping patches of  $64 \times 64$  pixels, which makes a total of 1088 samples for seventeen balanced classes.

The experiments were held using two different tessellations of the frequency domain. For the first one, five dyadic scales (the maximum starting from width one and covering all the image) and eight orientations were used. For the second one, eight constant-width frequency bands and eight orientations were considered. It has been introduced certain degree of overlapping between filters as recommended in [1]. Gaussian distributions are designed to overlap each other when achieving a value of 0.5. The three kind of features previously described have been tested: Gabor features using complex channels over RGB images, Gabor features over complex PCA-RGB channels, and color opponent features. As stated in previous section, only four orientations were considered for color opponent feature due to symmetry.

For each of the scales considered a classification experiment was held using only the features provided for that scale. In addition, an analysis of the combination of adjacent scales have been performed. In order to study the importance of low frequencies an ascending joining was performed, characterizing patch with the data provided by joined

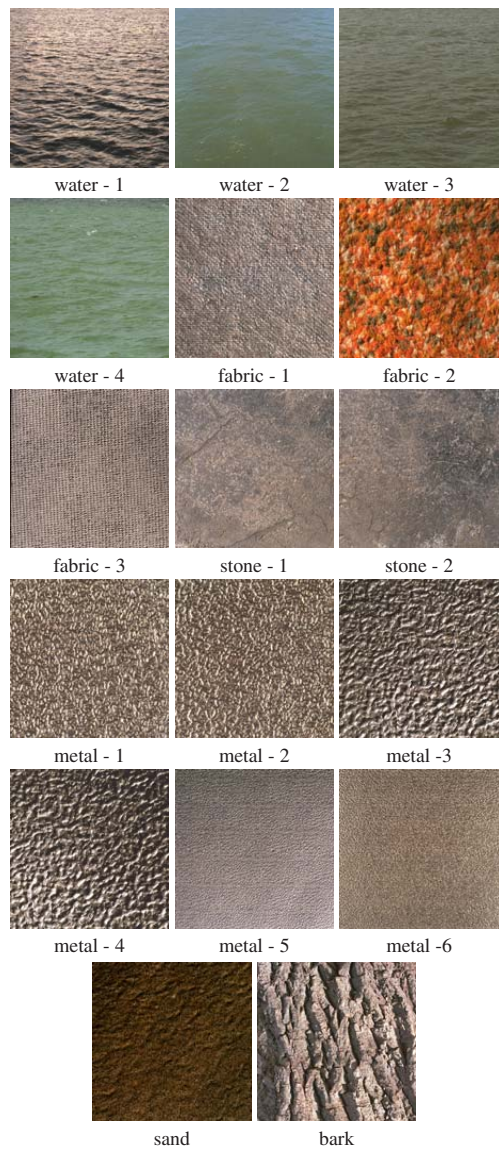


Fig. 1. Color textures used in the experimental campaign

ascendant scales. Similarly, the study of the high frequencies was carried out by a descendant joining. Also for medium frequencies, central scales are considered initially and adjacent lower and higher scales are joined gradually.

All texture patches characterized are later divided into sixteen separate sets keeping the a priori probability of each texture class. Therefore, no redundancies are introduced and each set is a representative set of the bigger original one. Eight classification attempts were carried out for each experiment with the k-nearest neighbor algorithm with  $k = 3$  and the mean of the correct classification rates of these attempts was taken as the final performance of the experiment. Each classification attempt uses one of these sets for training and another one as test set. Therefore, each set was never used twice in the same experiment.

#### 4 Experimental Results

Figure 2 shows the percentages of correct sample classification obtained for the experiments that used the dyadic filter banks whereas Figure 3 shows similar experimentation when the constant width filter banks were used instead.

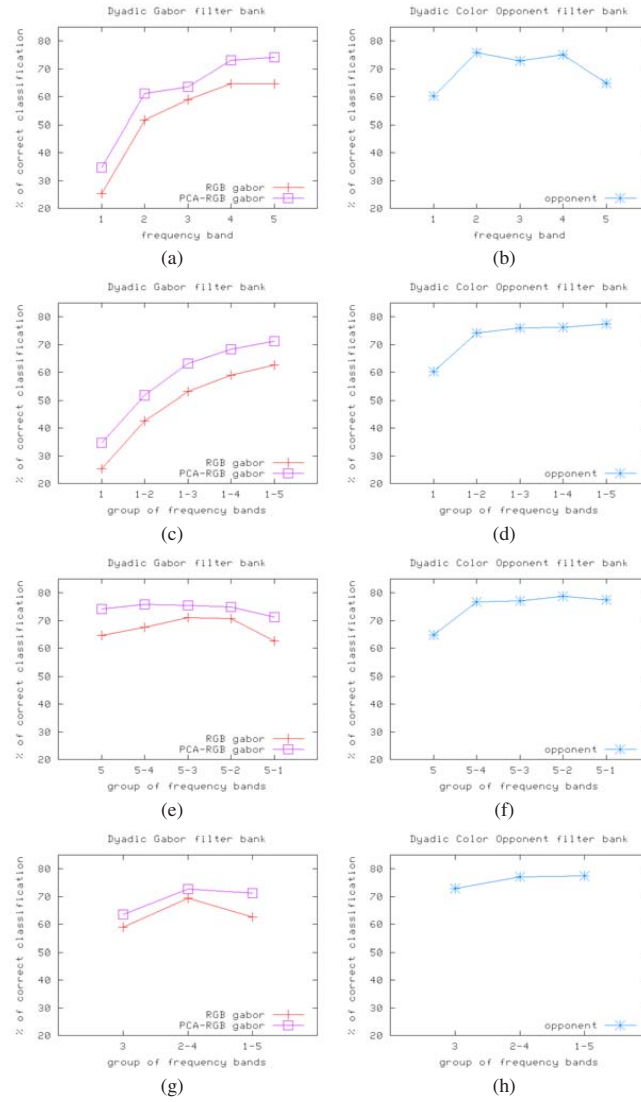
As it can be observed in both figures, the filter bank using the constant tessellation outperforms the dyadic one being in general more consistent. Briefly, the more detail is obtained from medium and high frequencies the best the texture is characterized. Note that a constant tessellation (Fig. 3) thoroughly analyzes medium and high scales which are claimed to contain discriminant information whereas dyadic does not. It can be observed in the graphs that, in general, the features derived from low scales do not help the characterization processes as the classification rates mainly decreases when they are considered.

By analyzing scales individually, Fig.2.(a-b) and 3.(a-b), the lower scale can never outperform the classification rates achieved by medium and high scales which, in some cases, achieve up to 75% by themselves.

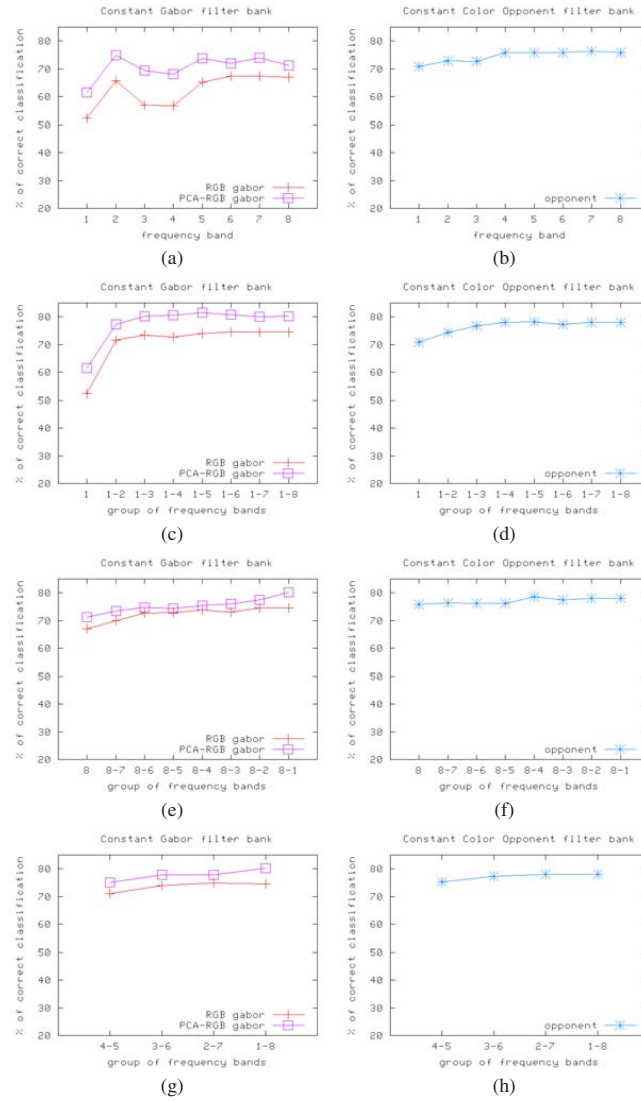
Regarding the dyadic tessellation, although scales 2-4 independently do not outperform the characterization using all scales together (Fig. 2a-b), their join performance does, Fig. 2g-h. This is because the scales by themselves do not cover the whole area containing outstanding information but their joining cover it all and consequently its performance reaches the maximum classification rates. It was expected that last scale outperformed the rest since it covers a larger frequential area. The ascendant joining presented in Fig. 2c-d shows a very poor performance for low frequencies and higher performances are not reached until medium frequencies take part of the characterization. Likewise, Fig. 2e-f enforce this conclusion showing high performances when taking medium frequencies into account. In a nutshell, when all (five) scales are used, the classification rates are better than the ones obtained using the medium scales independently. However, it is similar to the results obtained joining this three scales although having a more reduced number of features which proves that medium frequencies include the main discriminant textural information.

Note that graphs in Fig. 3 outperform those commented before. This is because medium and high frequencies are better analyzed in this case and this leads to a better texture characterization, improving the performance for all sort of joinings.

## Scale Analysis of Several Filter Banks for Color Texture Classification 515



**Fig. 2.** Pixel classification rates using the filter bank with dyadic tessellation. (a,c,e,g) Gabor features over complex planes (b,d,f,h) Opponent features (a,b) Individual scales (c,d) Ascendent join (e,f) Descendent join (g,h) Central join.



**Fig. 3.** Pixel classification rates using the filter bank with constant tessellation. (a,c,e,g) Gabor features over complex planes (b,d,f,h) Opponent features (a,b) Individual scales (c,d) Ascendent join (e,f) Descendent join (g,h) Central join.



Fig. 3g-h shows that no increase of the features may improve the characterization output as performance stays in the same values obtained using medium frequencies.

Last but not least, the comparison between the feature extraction methods suggests that opponent features perform slightly better than Gabor filters over complex bands using RGB channels. It seems that opponent features provides an efficient method of including inter-channel information while decreasing correlation among these channels. This points out that inter-channel interaction is also very important for characterization and color images should not be treated as a simple dimensional extension. For this reason, when PCA is considered before the application of Gabor filters over complex channels, their results outperform not only the classification rates obtained using the original RGB channels, but also the rates obtained using the color opponent features. It is important to bear in mind that, in this case, the number of features used to characterize each texture patch is significantly smaller than the number of color opponent features.

## 5 Conclusions

An analysis of the contribution of each scale to the characterization of color texture images has been performed. As it is known in the texture analysis field, medium and high frequencies play an essential role in texture characterization. Consequently, as has been shown, a constant tessellation of the frequency domain outperforms the traditional dyadic tessellation for microtexture characterization. For three different feature extraction methods, a thoroughly analysis of the contribution of each independent scale and the groups composed by low, medium or high frequencies has been carried out. Besides, a few scales could be considered in the feature extraction process providing by themselves very high classification rates with a reduced number of features. The experiments carried out have shown that the usage of PCA in RGB images before applying the Gabor filters over complex channels enhance the texture characterization significantly. Furthermore, these features outperformed the color opponent features even using a smaller number of features.

## Acknowledgment

This work has been partly supported by grant FPI PREDOC/2007/20 and project P1-1B2007-48 from Fundació Caixa Castelló-Bancaixa and projects CSD2007-00018 (Consolider Ingenio 2010) and AYA2008-05965-C04-04 from the Spanish Ministry of Science and Innovation.

## References

1. Bianconi, F., Fernández, A.: Evaluation of the effects of Gabor filter parametres on texture classification. *Patt. Recogn.* 40, 3325–3335 (2007)
2. Chang, T., Kuo, C.C.J.: Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. on Geoscience & Remote Sensing.* 441, 429–441 (1993)
3. Fogel, I., Sagi, D.: Gabor filters as texture discrimination. *Biological Cybernetics* 61, 103–113 (1989)

4. Grigorescu, S.E., Petkov, N., Kruizinga, P.: Comparison of Texture Features Based on Gabor Filters. *IEEE Trans. Image Processing* 11(10), 1160–1167 (2002)
5. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture Features for Image Classification. *IEEE Trans. Systems, Man, and Cybernetics* 3(6), 610–621 (1973)
6. Jaim, A., Healey, G.: A multiscale representation including opponenent color features for texture recognition. *IEEE Trans. Image Process.* 7, 124–128 (1998)
7. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on PAMI* 11, 674–693 (1989)
8. Ojala, T., Pietikainen, M., Maaenpaa, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
9. Petrou, M., García-Sevilla, P.: *Image Processing: Dealing with Texture*. John-Wiley and Sons, Dordrecht (2006)
10. Rajadell, O., García-Sevilla, P.: Influence of color spaces over texture characterization. *Research in Computing Science* 38, 273–281 (2008)
11. Randen, T., Hakon Huosy, J.: Filtering for Texture Classification: A Comparative Study. *IEEE Trans. Pattern Analysis and Machine Intelligence* 21(4), 291–310 (1999)
12. Varma, M., Zisserman, A.: Texture classification: Are filter banks necessary? In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 691–698 (2003)
13. VisTex Texture Database, MIT Media Lab (1995),  
<http://vismod.media.mit.edu/>

## Filter Banks for Hyperspectral Pixel Classification of Satellite Images

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

Depto. Lenguajes y Sistemas Informáticos  
Jaume I University, Campus Riu Sec s/n 12071 Castellón, Spain  
{orajadel,pgarcia,pla}@lsi.uji.es  
<http://www.vision.uji.es>

**Abstract.** Satellite hyperspectral imaging deals with heterogenous images containing different texture areas. Filter banks are frequently used to characterize textures in the image performing pixel classification. This filters are designed using different scales and orientations in order to cover all areas in the frequential domain. This work is aimed at studying the influence of the different scales used in the analysis, comparing texture analysis theory with hyperspectral imaging necessities. To pursue this, Gabor filters over complex planes and opponent features are taken into account and also compared in the feature extraction process.

### 1 Introduction

Nowadays imaging spectrometers are significantly increasing their spatial resolution. As their resolution increases, smaller areas are represented by each pixel in the images, encouraging the study of the relations of adjacent pixels (texture analysis) [9] [6]. However, not only the spatial resolution increases but also the spectral resolution. This entails dealing with a large number of spectral bands with highly correlated data [7].

Both dimensionality and texture analysis in hyperspectral imagery have been tackled from different points of view in literature. Several solutions to the dimensionality problem can be found, such as selection methods based on mathematical dimensionality reduction [10] or methods based on information theory which try to maximize the information provided by different sets of spectral bands [7].

Moving to texture analysis, literature survey provides us with a wide variety of well known texture analysis methods based on filtering [8] [4]. It is well known that, when dealing with microtextures, the most discriminant information falls in medium and high frequencies [1] [9]. It has been recently proposed that spatial/texture analysis may significantly improve the results in pixel classification tasks for satellite images using a very reduced number of spectral bands [11]. Therefore, it may be convenient to identify the influence of each frequency band separately in order to compare the textural information with the specific necessities of hyperspectral satellite imaging.

Besides, color opponent features were first introduced in color texture characterization with fairly good performance [3] and later extended to deal with multi-band texture images [4]. However, they have never been used to perform pixel classification tasks in satellite images. In this paper, we study several Gabor filter banks as well as multi-band opponent features for pixel classification tasks.

## 2 Filter Banks and Feature Extraction

Applying a filter over an image band provides a response for each pixel. If a filter bank is applied, a pixel can be characterized by means of the responses generated by all filters. It is possible to apply a filter in the space domain by a convolution or in the frequency domain by a product. In both cases, the response is the corresponding part of the original pixel value which responds to the filter applied [12].

When using filter banks, they are generally designed considering a dyadic tessellation of the frequency domain, that is, each frequency band (scale) considered is double the size of the previous one. It should not be ignored that this tessellation of the frequency domain thoroughly analyzes low frequencies giving less importance to medium and higher frequencies. Because the purpose of this work is to study the importance of texture in the pixel classification task, an alternative constant tessellation (given the same width to all frequency bands) is proposed in order to ensure an equal analysis of all frequencies.

### 2.1 Gabor Filters

Gabor filters consist essentially of sine and cosine functions modulated by a Gaussian envelope that achieve optimal joint localization in space and frequency. They can be defined by eq. (1) and (2) where  $m$  is the index for the scale,  $n$  for the orientation and  $u_m$  is the central frequency of the scale.

$$f_{mn}^{real}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_mx \cos \theta_n + u_my \sin \theta_n)) \quad (1)$$

$$f_{mn}^{imag}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \sin(2\pi(u_mx \cos \theta_n + u_my \sin \theta_n)) \quad (2)$$

If symmetrical filters are considered only the real part must be taken into account.

### 2.2 Gabor Filters over Complex Planes

Texture analysis in multi-channel images has been generally faced as a multi-dimensional extension of techniques designed for mono-channel images. In this way, images are decomposed into separated channels and the same feature extraction process is performed over each channel. This fails in capturing the interchannel properties of a multi-channel image.

To describe the inter-channel properties of textures we propose features obtained using Gabor filters over complex planes. This means that instead of using each spectral band individually, we take advantage of the complex definition and introduce the data of two spectral bands into one complex band, one as the real part and the other one as the imaginary part. In this way we involve pairs of bands in each characterization process, as it happens for the opponent features. As a result, for a cluster of spectral bands, we will consider all possible complex bands (pairs of bands). The Gabor filter bank will be applied over all complex bands as shown in eq. 3, where  $I^i(x, y)$  is the  $i^{th}$  spectral band.

$$h_{mn}^{ij}(x, y) = (I^i(x, y) + I^j(x, y)i) * f_{mn}(x, y) \quad (3)$$

The feature vector for each pixel in the image is composed of the response for that pixel to all filters in the filter bank, that is:

$$\psi_{x,y} = \{h_{mn}^{ij}(x,y)\}_{\forall i,j/i \neq j, \forall m,n} \quad (4)$$

The size of the feature vector varies with the number of complex bands. For each complex band, one feature is obtained for each filter applied what means that there will be as many features as filters for each complex band and as many complex bands as combinations without order nor repetition may be done with two bands in the cluster  $B$ . The total number of features is given by eq. 5 where  $M$  stands for the number of scales and  $N$  for the number of orientations.

$$size(\psi_{x,y}) = M \times N \times \binom{B}{2} \quad (5)$$

### 2.3 Opponent Features

Opponent features combine spatial information across spectral bands at different scales and are related to processes in human vision [3]. They are computed from Gabor filters as the difference of outputs of two different filters. The combination among filters are made for all pair of spectral bands  $i, j$  with  $i \neq j$  and  $|m - m'| \leq 1$ :

$$d_{mm'n}^{ij}(x,y) = h_{mn}^i(x,y) - h_{m'n}^j(x,y) \quad (6)$$

In this case, the feature vector for a pixel is the set of all opponent features for all spectral bands.

$$\varphi_{x,y} = \{d_{mm'n}^{ij}(x,y)\}_{\forall i,j/i \neq j, \forall m,m'/|m-m'| \leq 1, \forall n} \quad (7)$$

Hence, the size of the opponent feature vector also depends on the number of bands, scales, and orientations:

$$\begin{aligned} size(\varphi_{x,y}) &= \left(\binom{B}{2} \times M + B^2 \times (M-1)\right) \times N = \\ &= size(\psi_{x,y}) + B \times (B-1) \times (M-1) \times N \end{aligned} \quad (8)$$

Note that, in this case, the number of features is considerably increased.

## 3 Experimental Setup

The hyperspectral image database 92AV3C image has been used in the pixel classification experiments. It was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) [13]. The 20-m GSD data was acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the original 220 AVIRIS spectral bands our band selection method provides us with ten clusters of bands which are sets of bands that are intended to maximize the information provided [7]. The first cluster contains just one bands, the second contains two bands, and so on.

The experimental activity was held using two filter banks. For the first one, six dyadic scales (the maximum starting from width one and covering all the image) and four orientations were used. For the second one, eight constant frequency bands and four orientations were considered. It has been introduced certain degree of overlapping as recommended in [2]. Gaussian distributions are designed to overlap each other when achieving a value of 0.5.

For each of the scales a classification experiment was held using only the features provided for that scale. In addition, an analysis of the combination of adjacent scales have been performed. In order to study the importance of low frequencies an ascendent joining was performed, characterizing pixels with the data provided by joined ascendent scales. Similarly, the study of the high frequencies was carried out by a descendant joining. Also for medium frequencies, central scales are considered initially and adjacent lower and higher scales are joined gradually.

The pixels in the image database are divided in twenty non overlapping sets keeping the a priori probability of each class. Therefore, no redundancies are introduced and each set is a representative set of the bigger original one. Ten classification attempts were carried out for each experiment with the k-nearest neighbor algorithm with  $k = 3$  and the mean of the error rates of these attempts was taken as the final performance of the classifier. Each classification attempt uses one of these sets for training and another as test set. Therefore, each set was never used twice in the same experiment.

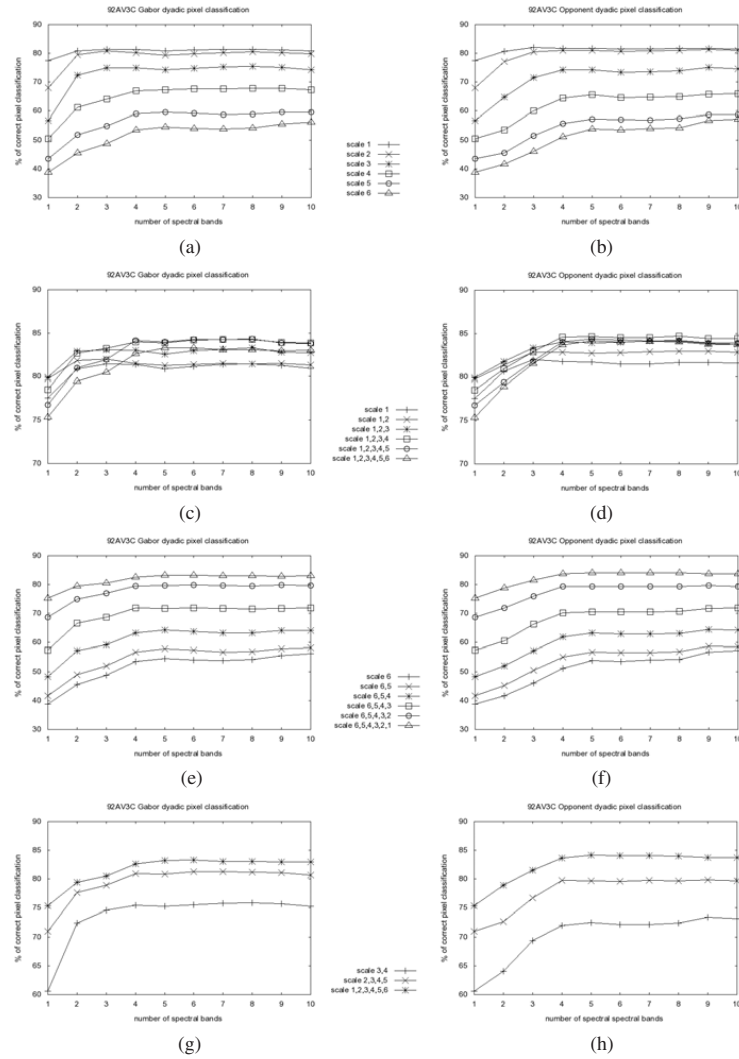
#### 4 Evaluation of the Results

Figure 1 shows the percentages of correct pixel classification obtained for the experiments that used the dyadic filter bank. Figure 2 shows similar results when the constant filter bank was used.

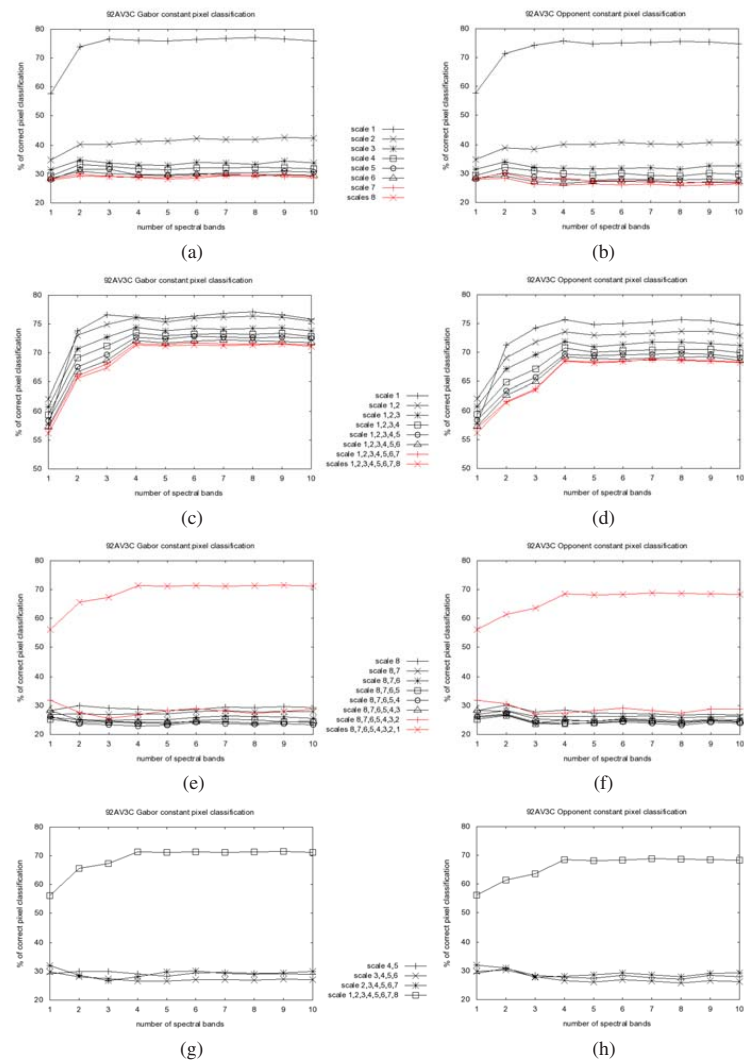
As it can be observed from both figures, when the characterization processes included all scales, the filter bank using the dyadic tessellation outperforms the constant one. It seems clear that the better the low frequencies are analyzed the better the pixels are characterized. This means that, for this sort of images, the texture information, although still helps in the characterization process, is significantly lower than the information contained in the low frequencies. It can be seen that no scale can ever outperform the classification rates achieved by scale one which achieve up to 81% by itself. In general, the more detail is obtained from low frequencies the best the image is characterized.

For the dyadic tessellation, although scales two and three do not outperform scale one when characterizing independently (Fig. 1a-b), their performance is considerably high. Because the first scales cover a very small part of the frequency domain, the characterization joining scales 1, 2 and 3 improve the pixel classification rates (Fig. 1c-d). In a nutshell, when all (six) scales are used, the classification rates are better than the ones obtained using just the first scale. However, it is worse than the results obtained for the first three scales although having a double number of features. The descendent and central joinings (Figs. 1e-f and 1g-h) clearly show that the performance increases significantly as features derived from lower frequencies are considered.

## Filter Banks for Hyperspectral Pixel Classification of Satellite Images 1043



**Fig. 1.** Pixel classification rates using the filter bank with dyadic tessellation. (a,c,e,g) Gabor features over complex planes (b,d,f,h) Opponent features (a,b) Individual scales (c,d) Ascendent join (e,f) Descendent join (g,h) Central join. Note the different ranges over the Y-axis in each graph.



**Fig. 2.** Pixel classification rates using the filter bank with constant tessellation. (a,c,e,g) Gabor features over complex planes (b,d,f,h) Opponent features (a,b) Individual scales (c,d) Ascendent join (e,f) Descendent join (g,h) Central join. Note the different ranges over the Y-axis in each graph.



Regarding the filter bank, using a constant tessellation (Fig. 2), the first scale is the only one containing discriminant information. This first scale is wide enough in this case to include the information of several scales of the dyadic tessellation. It is very clear from the graphs that the features derived from other scales do not help the characterization processes as the classification rates always decrease. It can be noticed that the best classification rates obtained for the dyadic tessellation is over 84% but is only about 77% for the constant tessellation.

Last but not least, the comparison between the feature extraction methods suggest that opponent features perform similarly to Gabor filters over complex planes. It seems that Gabor features provide better results when using a very small number of spectral bands whereas opponent features provide slightly higher classification rates when more spectral bands are used. Nevertheless, on the whole, the characterization with opponent features requires a larger number of features than Gabor filters, which may worsen performance if a large number of spectral bands is to be considered.

Briefly, spatial analysis between pixels improves hyperspectral satellite images characterization [11] but the nature of this kind of images, which are heterogeneous due to being composed of different homogeneous areas, made low frequencies very important for the characterization task, while texture information may help the process, but not significantly. Furthermore, including much more information but the provided by the low frequency analysis may even decrease the performance because of the so call Hughes phenomenon [5].

## 5 Conclusions

An analysis of the contribution of each scale to the characterization of hyperspectral images has been performed. As it is known in the texture analysis field, medium and high frequencies play an essential role in texture characterization. However, satellite images cannot be considered as pure texture images since the homogeneity of the different areas in the image is more important than the texture these areas may content. A thoroughly analysis of the contribution of each independent scale and the group composed by low, medium or high frequencies has been carried out. It has been shown that a detailed analysis of low frequencies helps the characterization improving performance. Also a few scales could be considered in the feature extraction process providing by themselves very high classification rates with a few number of features. The comparison between Gabor filters over complex plains and opponent features has shown that the classification rates obtained are almost identical in both cases. The main difference is the number of features required in each case, being much larger for the opponent features.

## Acknowledgment

This work has been partly supported by Fundació Caixa Castelló-Bancaixa through grant FPI PREDOC/2007/20 and project P1-1B2007-48, project CSD2007 00018 from Consolider Ingenio 2010, and project AYA2008-05965-C04-04 from Spanish CICYT.

## References

1. Chang, T., Kuo, C.C.J.: Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. on Geoscience & Remote Sensing* 2, 429–441 (1993)
2. Bianconi, F., Fernández, A.: Evaluation of the effects of Gabor filter parameters on texture classification. *Patt. Recogn.* 40, 3325–3335 (2007)
3. Jaim, A., Healey, G.: A multiscale representation including opponent color features for texture recognition. *IEEE Trans. Image Process.* 7, 124–128 (1998)
4. Shi, M., Healey, G.: Hyperspectral texture recognition using a multiscale opponent representation. *IEEE Trans. Geoscience and remote sensing* 41, 1090–1095 (2003)
5. Hughes, G.F.: On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* 14, 55–63 (1968)
6. Landgrebe, D.A.: *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, Hoboken (2003)
7. Martínez-Usó, A., Pla, F., Sotoca, J.M., García-Sevilla, P.: Clustering-based Hyperspectral Band selection using Information Measures. *IEEE Trans. on Geoscience & Remote Sensing* 45(12), 4158–4171 (2007)
8. Mercier, G., Lennon, M.: On the characterization of hyperspectral texture. *IEEE Trans. Inf. Theory* 14, 2584–2586 (2002)
9. Petrou, M., García-Sevilla, P.: *Image Processing: Dealing with Texture*. John Wiley and Sons, Chichester (2006)
10. Plaza, A., Martínez, P., Plaza, J., Pérez, R.: Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. *IEEE Trans. Geoscience and remote sensing* 43(3), 466–479 (2005)
11. Rajadell, O., García-Sevilla, P., Pla, F.: Textural features for hyperspectral pixel classification. In: Araujo, H., et al. (eds.) *IbPRIA 2009. LNCS*, vol. 5524, pp. 497–504. Springer, Heidelberg (2009)
12. Shaw, G., Manolakis, D.: Signal processing for hyperspectral image exploitation. *IEEE Signal Process. Mag.* 19(1), 12 (2002)
13. Vane, G., Green, R., Chrien, T., Enmark, H., Hansen, E., Porter, W.: The Airborne Visible Infrared Imaging Spectrometer *Remote Sens. Environ.* 44, 127–143 (1993)

# Spectral–Spatial Pixel Characterization Using Gabor Filters for Hyperspectral Image Classification

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

**Abstract**—This letter presents a spectral–spatial pixel characterization method for hyperspectral images. The characterization is based on textural features obtained using Gabor filters over a selected set of spectral bands. This scheme aims at improving land-use classification results, decreasing significantly the number of spectral bands needed in order to reduce the dimensionality of the task owing to an adequate description of the spatial characteristics of the image. This allows requiring less data and avoiding the curse of dimensionality. Very promising results are obtained which are similar to or better than previous classification results provided by other spectral–spatial methods but here also reducing the complexity using a reduced number of spectral bands.

**Index Terms**—Classification, hyperspectral, image segmentation, texture.

## I. INTRODUCTION

CURRENT hyperspectral sensors can have high spectral and spatial resolution. Some sensors can cover spectral resolutions higher than 10 nm, reaching 1 m per pixel for spatial resolution (e.g., some images provided by the ROSIS sensor). As a result, hyperspectral images are composed of a high number of correlated bands that may cause a dimensionality problem. When the spatial resolution was not so high, main efforts were focused at the classification stage. In particular, support vector machines (SVMs) proved to obtain good performances in this task [1]. With the increase of the spatial resolution, a joint spectral–spatial analysis was identified as a desired goal [2]. Spectral–spatial characterization aims at obtaining one feature vector for each pixel in the image based on the spectral measurements (spectral information) and a series of values extracted from spatial operations involving neighboring pixels (spatial information). However, as in all classification problems, it should not be forgotten that increasing the number of features used does not provide an endless improvement because of the well-known curse-of-dimensionality problem [3].

Nowadays, a wide range of techniques is used to include spatial information into the image characterization, such as morphological profiles [4] or Markov fields [5]. However, these methods introduce a scale selection problem. Recently, several proposals have been developed to face the overseg-

mentation problem and the scalability with very good results. Tarabalka *et al.* [6] presented a spectral–spatial classification scheme that consists of a pixelwise classification and a partitioning clustering by a majority vote with adaptive neighborhoods. The result is a segmentation map that needs a spatial postregularization to reduce the noise. This provides more homogeneous regions than a simple pixelwise classification process, but it is not yet suitable for images containing small classes since they may be missed. The same problem is observed in [7], where an extension of the watershed segmentation algorithm for hyperspectral images was presented in order to define the spatial structures. To deal with the segmentation of small regions, the same authors suggested in [8] to select the most reliable pixels from a pixelwise classification as markers to be used in a minimum spanning forest grown obtaining a spectral–spatial classification map refined afterward by majority voting within the spatially connected regions.

The characterization of spatial structures in an image has been studied in detail when dealing with the analysis of visual textures [9]. However, most of these methods were developed mainly for gray-level images, and their extension for multichannel images has been generally faced as a multidimensional extension of the monochannel techniques. Jaim and Healey [10] made one of the first proposals on how to use spatial information across spectral bands using Gabor filters. Opponent features were first described for color images [10] and extended to be used over multichannel images [11]. They combine spatial information across spectral bands at different scales by combining the responses of the filters applied separately to each channel. Lately, they also used 3-D Gabor filter banks [12]. However, all these methods have been always applied to patches of stationary textures, and no analysis has been done about the characterization of individual pixels which allows segmentation of images using this spatial information.

In order to segment and classify hyperspectral land cover images, we classify individual pixels to get a classification map following a late trend [5], [6], [8]. This task has already been faced using a large amount of data. However, when devices improve, dealing with an increasing amount of data also increases the risk of reaching the accuracy ceiling. Thus, we also aim at using a very small number of features to obtain the same or even better results found in the literature, leaving then room for adding new features that may improve the classification. To pursue this objective, a band selection method will be first used over the whole set of bands provided by the spectrometer. Then, the pixel characterization methods will be applied over the selected spectral bands. Three different pixel characterization methods based on Gabor filters will be used here.

The rest of this letter is organized as follows. First, filter bank characterization methods are introduced in Section II, Gabor

Manuscript received April 1, 2012; revised September 9, 2012; accepted October 9, 2012. Date of publication December 19, 2012; date of current version February 20, 2013. This work was supported in part by FPI under PREDOC/2007/20 and in part by the Generalitat Valenciana under Projects CSD2007-00018 (Consolider Ingenio 2010), AYA2008-05965-C04-04, and PROMETEO/2010/028.

The authors are with the Institute of New Imaging Technologies, University Jaume I, 12071 Castellón, Spain (e-mail: orajadel@uji.es; pgarcia@uji.es; pla@uji.es).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.  
Digital Object Identifier 10.1109/LGRS.2012.2226426

filters have been used for texture characterization, and we will propose their use for pixel classification in two different ways. Moreover, we will adapt a method from the literature to perform pixel classification and compare all of them. In Section III, we present the database used in this letter followed by the classification setup that will be further used, a comparison between the three characterization methods, and a study of the relation between the characterization and the scales of the filters within the filter bank. The supervised segmentation results are presented in Section IV as images, also providing the per-class accuracies. Eventually, conclusions can be found in Section V.

## II. FILTER BANK CHARACTERIZATION METHODS

Several features have been suggested in the literature for the description of spatial (or texture) information (see [9] for a survey). In this letter, features are obtained by filtering the input image with a set of filters (filter bank). The vector of features per pixel corresponds to all the responses of the pixel to the filter bank.

For an image of  $B$  bands, let  $I^i$  be the  $i$ th band. Let  $f_k$  be the  $k$ th filter in the filter bank  $F$ . The response to the filter when applied over the  $i$ th band is given by  $h_k^i = I^i * f_k$ , where  $*$  stands for the convolution operator.

We chose to use a Gabor filter bank. This is a set of Gabor filters of  $M$  different scales (spatial frequencies) and  $N$  orientations designed to cover the frequency domain

$$F = \{f_{m,n}\}_{m=1,n=1}^{M,N}. \quad (1)$$

They consist of sine and cosine functions modulated by a Gaussian envelope that achieve optimal joint localization in space and frequency [13]. They can be defined by

$$f_{mn}^{\text{real}}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_m x \cos \theta_n + u_m y \sin \theta_n)) \quad (2)$$

$$f_{mn}^{\text{imag}}(x, y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma_m^2}\right\} \times \sin(2\pi(u_m x \cos \theta_n + u_m y \sin \theta_n)) \quad (3)$$

where  $m$  is the index for the scale,  $n$  is the index for the orientation, and  $u_m$  is the central frequency of the scale [14].

For real signal values, the outputs for orientations  $\theta_n$  and  $\theta_n + \pi$  will be complex conjugates. These pairs of filters are usually joined together, canceling in this way the imaginary parts of the outputs and dealing only with real value outputs.

### A. Opponent Features

Opponent features [10] use Gabor filters and combine the filtered results (spatial information) across spectral bands at different scales. According to the authors, this is related to processes in human vision. They are computed from the responses to Gabor filters as the difference of responses between two different filters. In other words, the spectral bands are first individually filtered, and their responses are combined afterward to obtain the opponent features aiming at introducing interchannel information into the characterization process. The combination among responses [11] is made for all pairs of

spectral bands  $i, j$ , with  $i > j$ , and two scales  $m$  and  $m'$ , such that  $0 \leq (m - m') \leq 1$ , as follows:

$$d_{mm'n}^{ij} = h_{mn}^i - h_{m'n}^j. \quad (4)$$

In our case, instead of computing the energies for whole image patches, a feature vector for each individual pixel is obtained as the set of all opponent features computed for it. In this way, we obtain opponent features for each individual pixel by applying the filter bank only once over the whole image. If a texture patch was considered around each pixel in the image, the filter bank must be applied over each patch. As each pixel will belong to several patches, it will be repeatedly analyzed. In this way, we expect to obtain results that are similar but reducing the computational effort required.

### B. Gabor Filters Over Individual Bands

We propose a simplified version where each spectral band is analyzed separately and each pixel is characterized with the responses to the filter bank used. This will result in a smaller number of features per pixel keeping the spatial information but missing the interchannel information.

When the whole filter bank is applied, the feature vectors for the pixels in the image will be obtained by simply taking the responses to all filters for all bands

$$\psi = \{h_k^i\}_{k=1,i=1}^{K,B}. \quad (5)$$

In this way, hyperspectral images will be simply decomposed into separated bands, and the same feature extraction process will be performed over each band. By filtering with such a filter bank, the response of one pixel to each filter is a decomposition of the spectral measurement in the amounts corresponding to each spatial frequency range and orientation used to define the filter bank.

### C. Gabor Filters Over Complex Bands

Filtering each band individually misses the interchannel information proposed by Jaim and Healey [10]. In order to test its significance, we propose a variation of the characterization method described previously that introduces interchannel data. To pursue this, two real bands are merged into one complex band, one as the real part and the other one as the imaginary part. Now, each Gabor filter will be applied over a complex band as follows:

$$h_{mn}^{ij} = (I^i + I^j i) * f_{m,n} \quad (6)$$

with  $i = \sqrt{-1}$ , where  $I^i$  and  $I^j$  are the  $i$ th and  $j$ th spectral bands, respectively.

All pairs of spectral bands will be considered and filtered. Interchannel information is included here because two spectral bands are filtered at once, so the response to the filters combines information from these two bands. Since the bands to be analyzed are no longer real, now, filters for orientations  $\theta_n$  and  $\theta_n + \pi$  will provide different outputs, and therefore, they are not joined together. As a consequence, the number of orientations will be the double of the number used for individual bands. Note also that, now, each output will be a complex number that will be represented using two real values while, in previous cases, each output was represented by just

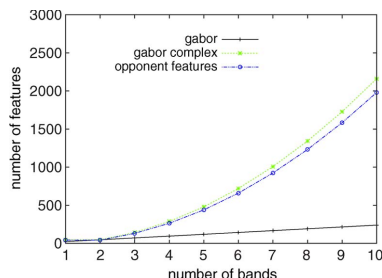


Fig. 1. Number of features per method versus the number of spectral bands. one real number. This doubles the number of features required in this case.

### III. CLASSIFICATION EXPERIMENTS

The scheme here proposed combines a band selection method with the spectral-spatial pixel characterization methods previously proposed. Among the different band selection methods, WaLuMI [15] has been chosen for preserving the original bands, providing as output a subset of them. It is based on a clustering of bands that pursues, as a whole, to maximize the mutual information among bands in each cluster and to minimize the intercluster correlation. However, any other band selection method that fulfills similar criteria can be used instead. In this section, all classification experiments are tested over the Indian Pine hyperspectral data set [Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)]. Two classifiers, SVM with a third-order polynomial kernel and a three-nearest neighbor (3-NN) classifier, are used.

#### A. Data Set

Hyperspectral image 92AV3C was provided by the spectrometer AVIRIS and acquired over the Indian Pine test site in Northwestern Indiana in 1992. The image has a spatial dimension of  $145 \times 145$  pixels. The spatial resolution is 20 m per pixel. Spectral coverage ranges from 0.38 to 2.50  $\mu\text{m}$  with 220 spectral bands. Classes range from 20 to 2468 pixels. Due to the small size of some classes, this database is suitable for testing if the proposed methods can also succeed at the classification of small areas which are often missed in highly unbalanced data sets.

#### B. Experiment Setup

For the characterization of the data, a Gabor filter bank is designed with four orientations and six scales. The four orientations ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$ ) are the minimum numbers of orientations recommended to get textural information [9]. Gabor filter scales are chosen to be dyadic, with the first scale having a width of one. Hence, given the size of the image used, the maximum number of scales is  $M = 6$ . Moreover, Gabor filters were designed to overlap each other when achieving a value of 0.5, following the recommendation in [16]. The filter bank is applied according to one of the methods defined in the previous section, and each pixel is characterized with the responses to it. This leads to a different number of features per pixel regarding the method used (see Fig. 1). This is an

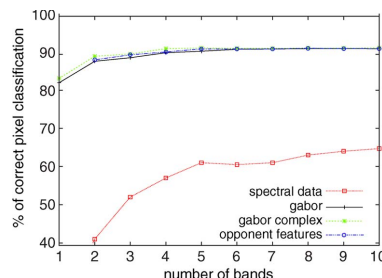


Fig. 2. Pixel classification rates for the proposed characterization methods over the AVIRIS database using an SVM classifier.

important issue because of the so-called Hughes phenomenon [3], which also leads to the fact that increasing the number of dimensions does not necessarily lead to an improvement.

For the classification experiments, the labeled pixels in the image database were divided into 20 nonoverlapping sets keeping the *a priori* probability of each class. Therefore, no redundancies were introduced. Ten classification attempts were carried out, and the mean of the error rates of these attempts was taken as the performance. For each attempt, one set was used for training and another was used for testing, and sets were never used twice. This methodology was already used in [15] and [17] in order to increase the statistical independence among the classification attempts.

#### C. Comparison of the Characterization Methods Proposed

In this section, the different characterization methods described in Section II are compared. The settings for these experiments are the ones described in Section III-B. The value of  $B$  (number of spectral bands) varies from one to ten in each experiment. The set of bands is provided by the WaLuMI algorithm.

The classification results using an SVM can be found in Fig. 2. The results using only spectral information were also included as a baseline reference. In all cases, the mean rate of ten experiments is shown. The variance between experiments was really small (less than 3%).

All spectral-spatial features clearly outperformed the spectral features. Also, we can see that there is almost no difference between the three spectral-spatial methods considered. Experiments using Gabor filters over texture patches around each pixel were also carried out, providing similar classification rates. This means that the spatial information is much more important than the interchannel information for the appropriate characterization of the pixels in the image. It is important to note that the initial information used in all experiments is exactly the same because the spatial features are directly computed from the spectral data.

The results obtained with the  $k$ -nearest neighbor classifier are slightly lower than the ones obtained with SVM (an average of 2% lower) with a small difference of 1.5% in favor of Gabor filters over complex bands.

Thus, we can conclude that spatial information improves the classification but the addition of interchannel information is not relevant enough and does not justify the increase in the dimension of the classification space. Considering this conclusion, for

the next series of experiments, we suggest the use of Gabor filters over individual bands.

For the number of bands, observe that, after  $B = 3$ , no significant improvement is achieved when increasing the number of bands. An experiment using all bands available ( $B = 220$ ) was performed with results of 87% using the SVM and 86.6% for the 3-NN classifier, which are below the maximum result shown in Fig. 2. Although  $B$  is a parameter to set for the process, the performance usually reaches a maximum, and adding more bands does not improve the classification results. Hence, the selection of  $B$  is not critical as long as we choose a value greater than the one needed to reach the flat zone of the learning curve.

#### D. Scale Analysis

In a Gabor filter bank, those filters with different orientations and the same scale provide information corresponding to the same range of spatial frequencies. It is known that different frequencies provide a different analysis of the scenario, for example, high frequencies contain most of the noise present in the image. The following experiment is a classification using solely the features obtained from each set of filters with the same scale but with different orientations. These results are shown in Fig. 3 (left). The settings for the classification experiments are the same as those in Section III-C, except for the scales used.

Note that, the lower the scale, the better the result. This was expected because most of the areas to classify are quite homogeneous.

As a further analysis, we also run an experiment performing a progressive combination of features from different scales. First, only features using the filters that are defined with the first range of spatial frequencies are taken. In each step, the features from the following scale are combined with the previous features by adding one scale at each step until covering the whole set of scales. These classification results are shown in Fig. 3 (right). Observe that, when we join the features of the first two scales, the classification rate improves. When adding the first, second, and third scales, the results are quite similar. However, when adding more than three scales, the results progressively worsen. Recall that the higher scales may mainly contain noise and they do not help in the characterization of the pixels. This highlights the fact that the discriminant piece of information for this sort of images is in the first scales because they contain well-defined areas of low spatial frequencies.

#### IV. SEGMENTATION EXPERIMENTS

To get a supervised segmentation from the pixelwise classification, we split our data in a set of samples with known labels and a test set to be classified. The resulting labels create a classification map. Unlike the previous experiments, the set of labeled pixels is here directly split in two. Five percent of the samples from the whole data set, keeping the *a priori* probabilities, form the training set, and the rest form the test set. Again, results using an SVM classifier are shown.

The classification results in Fig. 2 show that the improvement has a maximum. Because any value of  $B$  over three will provide a similar result, raising the number of features (dimensions) in this problem is not convenient. Hence, the set of bands for  $B = 3$  was selected using the WaLuMI algorithm. Furthermore,

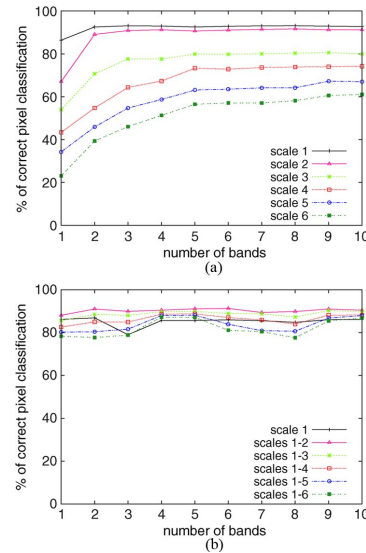


Fig. 3. Pixel classification rates for the AVIRIS data set using spatial features derived from Gabor filters and an SVM classifier. (Left) Analysis of individual scales. (Right) Joining of features from consecutive ascending scales.

as seen in Section III-D, adding features from more scales is neither improving the characterization. For that reason, we chose to perform this experiment with the features coming from  $B = 3$  and combining the features from the first two scales. We chose to reduce the number of features to show that 24 features can provide a result as good as or even better than a much higher number of dimensions.

The global classification accuracy obtained was 92.99% using the SVM (note that the result using the 3-NN classifier was the same). This result is slightly higher than the ones in [7] and [8], where the same problem for the AVIRIS data set was addressed, obtaining 91.80% of correct classification. Moreover, in these cited works, a fixed number of samples per class were picked as training set; thus, the *a priori* probabilities were not kept, small classes were overrepresented in the training, and all spectral bands were used there. Therefore, the results presented here have been obtained in more realistic conditions, taking into account that real unbalanced data are a harder classification problem.

The producer's accuracy per class for the AVIRIS data set is shown in Table I. Notice that 7 of the 16 classes are usually ignored in this sort of experiments because they contain a very small number of pixels [1]. However, we include them in our experiments, and the results are fairly good, considering the difficulties when treating with such unbalanced classification problem. For example, the class representing Oats has only 20 pixels, and only one pixel was used for training. Therefore, an important amount of classification errors is expected. Nonetheless, it is remarkable that other small areas corresponding to Alfalfa, Bldg-Grass-Trees-Drives, Grass/pasture-mowed, Corn, and Wheat were fairly well classified. We can also see the same results in the image in Fig. 4, where the errors are represented in white.



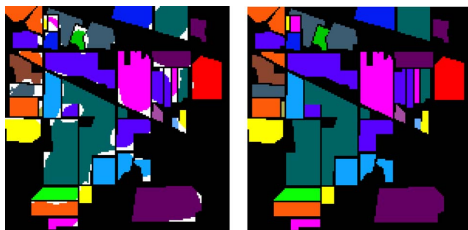


Fig. 4. (Left) Classification results for the AVIRIS data set using 24 features (four orientations, two scales, and three spectral bands). Overall accuracy: 84.836. Kappa coefficient: 0.791. (Right) Ground truth.

TABLE I  
PRODUCER'S ACCURACY PER CLASS FOR THE AVIRIS DATA SET  
USING 24 FEATURES (FOUR ORIENTATIONS, TWO SCALES,  
AND THREE SPECTRAL BANDS)

class	training/total	Per class accuracy (%)
Stone-steel towers	4/95	76.92
Hay-windrowed	25/489	99.14
Corn-min till	42/834	96.59
Soybeans-no till	48/968	89.23
Alfalfa	2/54	100.0
Soybeans-clean till	30/614	87.32
Grass/pasture	25/497	90.04
Woods	65/1294	95.77
Bldg-Grass-Tree-Drives	19/380	97.22
Grass/pasture-mowed	2/26	91.66
Corn	11/234	92.82
Oats	1/20	52.63
Corn-no till	72/1434	92.51
Soybeans-min till	124/2468	91.93
Grass/trees	37/747	94.92
Wheat	10/212	99.50
Overall accuracy		92.99
kappa		0.92

## V. CONCLUSION

A hyperspectral pixel classification scheme that combines a band selection procedure with a spatial feature extraction process has been presented. The increase of the spatial resolution in hyperspectral sensors encouraged this idea. It has been experimentally proven that the proposed scheme provides better classification rates than other state-of-the-art spectral-spatial methods. Furthermore, the approach presented here uses a reduced set of selected spectral bands, simplifying the representation while keeping the classification rates with respect to other approaches. This is important in order to avoid the problems caused by the curse of dimensionality and also because it leaves room for other features to be used to improve the characterization.

Three spatial features have been suggested for the characterization of individual pixels, all of them based on features derived from Gabor filters. We have shown that the spatial information provides an appropriate characterization of the pixels for classification tasks. These features lead to good classification rates. We have also shown that the spatial information influences the characterization process much more than the interchannel information. No big differences have been found between the three sorts of spatial features analyzed, although they have big differences in the number of features used to describe each pixel, with the method proposed by applying Gabor filters over

individual bands being the most appropriate because of its simplicity and smaller dimensionality.

We have also studied the influence of the different scales in the feature extraction process and found that, when only areas of low spatial frequencies compose the image, the first scales provide the best characterization and the addition of the last scales tends to worsen the classification results. However, if we have to deal with nonhomogeneous regions, the use of the medium scales may improve the characterization.

In the segmentation experiments, we found that most of the misclassified pixels fall in the borders of the labeled regions where the spatial features can be confused due to the background information or due to the transitions between different classes in the image plane. However, the segmentation of the inner part of the regions was always remarkably homogeneous, despite the fact that no further spatial regularization is applied to the pixel-based classification proposed.

## REFERENCES

- [1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image proc.," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, Sep. 2009.
- [2] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.
- [3] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
- [4] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 3804–3814, Nov. 2008.
- [5] J. Li, J. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.
- [6] Y. Tarabalka, J. Chanussot, and J. Benediktsson, "Spectral-spatial classification of hyperspectral imagery based on partitioned clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.
- [7] Y. Tarabalka, J. Chanussot, and J. Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pattern Recognit.*, vol. 43, no. 7, pp. 2367–2379, Jul. 2010.
- [8] Y. Tarabalka, J. Chanussot, and J. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 40, no. 5, pp. 1267–1279, Oct. 2010.
- [9] M. Petrou and P. García-Sevilla, *Image Processing: Dealing With Texture*. Hoboken, NJ: Wiley, 2006.
- [10] A. Jaim and G. Healey, "A multiscale representation including opponent color features for texture recognition," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 124–128, Jan. 1998.
- [11] M. Shi and G. Healey, "Hyperspectral texture recognition using a multiscale opponent representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1090–1095, May 2003.
- [12] T. Bau, S. Sarkar, and G. Healey, "Hyperspectral region classification using three-dimensional Gabor filterbank," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3457–3464, Sep. 2010.
- [13] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biol. Cybern.*, vol. 61, no. 2, pp. 103–113, 1989.
- [14] A. Jain and F. Farrokhi, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, 1991.
- [15] A. Martínez-Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.
- [16] F. Bianconi and A. Fernández, "Evaluation of the effects of Gabor filter parameters on texture classification," *Pattern Recognit.*, vol. 40, no. 12, pp. 3325–3335, Dec. 2007.
- [17] J. S. Sánchez, R. A. Mollineda, and J. M. Sotoca, "An analysis of how training data complexity affects the nearest neighbor classifiers," *Pattern Anal. Appl.*, vol. 10, no. 3, pp. 189–201, Jul. 2007.

## SEMI-SUPERVISED HYPERSPECTRAL PIXEL CLASSIFICATION USING INTERACTIVE LABELING

*Olga Rajadell, Pedro García-Sevilla*

University Jaume I  
 Depto. Lenguajes y Sistemas Informáticos  
 Institute of New Imaging Technologies  
 Castellón, Spain  
 orajadel,pgarcia@lsi.uji.es

*V.C. Dinh<sup>+</sup>, R. P. W. Duin<sup>+</sup>*

<sup>+</sup>Delft University of Technology  
 Department of Mediamatics  
 Delft, The Netherlands

\*Carinthian Tech Research AG  
 Department of Spectral Imaging  
 Villach, Austria  
 v.c.dinh@tudelft.nl, r.duin@ieee.org

### ABSTRACT

A semi-supervised pixel classification scheme for hyperspectral satellite images is presented. The scheme includes a previous band selection step followed by a clustering process to select modes of interest that will be labeled by an expert. Then pixel classification is performed resulting in a segmentation and classification of the fields appearing in the image. Thanks to the previous clustering step the most suitable pixels are automatically selected to build the classifier. This reduces the expert effort required since less pixels need to be labeled. However pixel classification accuracy obtained outperforms the results of a random selection scheme where many more pixels were labeled.

**Index Terms**— Pixel classification, hyperspectral imaging, semi-supervised classification, mode seek, spectral/spatial features.

### I. INTRODUCTION

Segmentation is a noted un-supervised issue in image processing research. Lately, this task has also been faced as a semi-supervised task in which experts provide labeled samples that the system can use to classify the pixels as well as to segment the image. To this end pixel classification is widely used but results may still need of additional information or process. In this direction, authors have tried to describe the neighborhood of the pixel using spectral/spatial features [1]. Other methods used MRF [2] suffering from the problem of setting a fixed shape. In [3] an adaptive neighborhood was defined to face this problem. Another popular strategy is defining a classification scheme that introduced a previous segmentation task [4] or a post-process

improvement [5]. But in all cases training sets are picked randomly over the dataset. It is always a drawback to reduce the size of the training set since randomly distributed pixels can lie in non interesting areas and consequently classes can be missed. On the contrary, the expert action is expected to be minimized in the labeling of the training samples. In this scenario the most interesting samples from the system point of view should be provided to the expert instead of the randomly selected ones. Tarabalka et al. introduced this idea in [6] focusing in the phase after pixel classification. This paper introduces a semi-supervised classification scheme aimed at decreasing the training samples before the classification task is performed.

Clustering algorithms analyze the feature space in order to group samples around a representant called mode. Thanks to nonparametric clustering techniques a feature space can be analyzed finding their modes in a non-supervised way. In this paper the random selection of samples required to train the classifier is suggested to be changed for the modes resulting of a clustering process of the samples. This non-supervised selection makes training samples suitable for posterior non-linear classification using a k-nearest neighbor rule.

The chosen clustering method is explained in Section II. Afterwards, the feature extraction and the semi-supervised classification scheme are presented in Section III and Section IV respectively. The database will be described in Section V. Later, in Section VI, the experiments will be presented and discussed. Conclusions on the whole paper can be found in Section VII.

### II. MODE SEEK CLUSTERING

Given a hyperspectral image, all pixels can be considered as samples which are characterized by their corresponding feature vectors (spectral curve). The set of features defined is called the feature space and samples (pixels) are represented as points in that multi-dimensional space. A clustering

THANKS TO FUNDACIÓ CAIXA CASTELLÓ-BANCAIXA FOR FUNDING BY GRANT FPI PREDOC/2007/20. ALSO TO THE SPANISH MINISTRY OF SCIENCE AND INNOVATION FOR SUPPORTING IN PROJECTS CSD2007-00018 (CONSOLIDER INGENIO 2010), AYA2008-05965-C04-04 AND MTM2009-14500-C02-02



method groups similar objects (samples) in sets that are called clusters. The similarity measure is defined by the clustering algorithm used. A crucial problem lies in finding a good distance measure between the objects represented by these feature vectors. Many clustering algorithms are well known. Among them, k-means is a widely used technique due to its ease of programming and good performance. However, k-means suffers from several drawbacks; it is sensitive to initial conditions, it does not remove undesirable features for clustering, and it is optimal only for hyper-spherical clusters. Furthermore, its complexity can be impractical for large datasets [7]. For such reasons a  $KNN$  mode-seeking method will be used in this paper. It selects a number of modes which is controlled by the neighborhood parameter ( $s$ ). For each class object  $x_j$ , the method seeks the dissimilarity to its  $s^{th}$  neighbors. Then, for the  $s$  neighbors of  $x_j$ , the dissimilarities to their  $s^{th}$  neighbors are also computed. If the dissimilarity of  $x_j$  to its  $s^{th}$  neighbor is minimum compared to those of its  $s$  neighbors, it is selected as prototype [8]. Note that the  $s$  parameter only influences the scheme in a way that the bigger it is the less clusters the method will get since more samples will be group in the same cluster, that is, less modes will be selected as a result.

### III. SPECTRAL/SPATIAL FEATURE EXTRACTION

Pixel characterization aims at obtaining one feature vector for each pixel to be used in a pixel classification task in a multidimensional space. When only spectral data is used the feature vector for every pixel is defined as the spectral curve provided by the sensor.

In order to describe the context of a pixel several features have been suggested in the literature [9]. In this paper Gabor filtering will be used as suggested in [1]. In this case, features are obtained by filtering the input image with a set of filters. The set of outputs obtained for each pixel in the image forms its feature vector. In this case, the filter bank is defined to be a set of two-dimensional Gabor filters. Each Gabor filter is characterized by a preferred orientation and a preferred spatial frequency (scale) and consist of sine and cosine functions modulated by a Gaussian envelope.

### IV. SEMI-SUPERVISED CLASSIFICATION

Here the proposed semi-supervised pixel classification scheme is presented. The scheme proceeds as follows:

- 1) In order to reduce the number of spectral bands to be used, a set of spectral bands, given a desired number, is selected by using the band selection method proposed in [10].
- 2) A Clustering procedure is applied over the selected spectral bands. An improvement in the clustering process is included by adding as features the spatial coordinates of each pixel in the image. This provides a spatial component very suitable for clustering since it is based in distances between samples.

- 3) The modes resulting of the previous step define the training set for the next step. The expert is involved in this point by providing the corresponding labels of the selected samples. Here the expert is simulated by checking the labels in the ground truth provided for only those samples.
- 4) A  $KNN$  classifier with  $k = 1$  is build with the training set defined above. Note that in this point the spatial coordinates are dismissed as features. Regarding the clustering step it is always performed over the spectral domain but, once the modes are obtained, the features to be used for the classification step can be the same or changed. In this paper classification step changing the space to spectral/spatial features is also tested.

The parameter  $s$  of the clustering algorithm can be tuned to obtained a higher or lower number of interesting points to be labeled. The increase of this parameter is inverse to the number of modes found. As it will be seen, the number of modes has a direct impact on the performance of the classification but still the results are better than the ones obtained using a random selection.

### V. DATASET

A widely used hyper-spectral database has been used in our experiments. Hyper-spectral image data 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR) [11]. The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. In it, three different growing states of soya can be found, together with other three different growing states of corn. Woods, pasture and trees are the bigger classes in terms of number of samples (pixels). Smaller classes can be also found such as steel towers, hay-windrowed, alfalfa, drives, oats, grass and wheat.

### VI. EXPERIMENTS, RESULTS AND EVALUATION

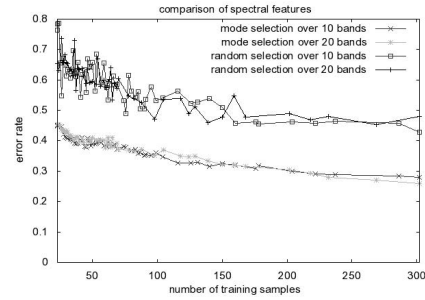
In Figures 1, 2 the performance of the semi-supervised classification scheme is compared with the traditional random selection and classification process. Results are shown as learning curves where error rate is represented as a function of the number of samples used for training. In Figure 1 learning curves for different number of spectral bands are presented together with the corresponding learning curve when the same amount of pixels are selected at random. It is noticeable that in all cases, when selecting the training set, the classification rate outperforms the result when it is picked at random. The gain reaches 0.3 when a smaller training set is used and decrease to 0.15 when the training set grows, obviously because when the size of the training set grows,

random selection has more chances to select samples from all different areas. Also, note that no advantage is obtained in involving a higher number of spectral bands in the process. If the number of spectral bands used to performed the clustering step is fixed to 10, similar conclusions can be obtained from Figure 2 where spectral/spatial features are used in this case. Using more than 3 bands leads to higher computational complexity with no performance increase. As a summary also the difference in the error rate between using 10 spectral and 24 spectral/spatial features derived from 3 bands for classification can be observed in Figure 3. Again, in both cases the error rate obtained using the random selection stays over the classification using the mode selection method. It is remarkable that both kind of features start around the same rate but the difference is quickly introduced when more samples are included and the error rate when using spectral/spatial features decrease considerably.

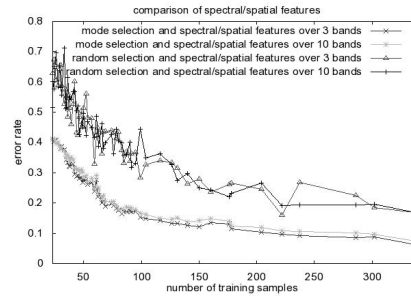
Showing the results over the image ground-truth, Figures 4 and 5 show results using 24 spectral/spatial features derived from 3 bands for the classifying step, when 23 and 104 training samples are selected respectively. In (a) misclassified pixels are represented in white color whereas the rest of the image represent well classified ones and (b) training pixels are presented in white over the ground-truth of the image. In both images background is the black area surrounded the classes and it is considered a non interesting heterogeneous area. It is very noticeable that small classes are missed in the mode selection when only 23 modes were found. That means that clustering method cannot detect those areas as independent ones. As a consequence of having no training sample available for that class, classification dismisses it all. As it can be expected, the smaller the number of clusters is, the higher number of small classes are missed. Nevertheless, where a sample is selected, a big area is well classified due to the usage of spectral/spatial features. Figure 4 stands for an error rate of 0.41 using only 23 samples as training set. Observe that only samples from 10 different classes are selected leading to miss 6 classes. However in Figure 5, using 104 training samples, the number of modes increases, 15 classes are included in the training set and the error rate decreases to 0.147.

These results may not seem significant in terms of figures. In [12] classification rates reached 95% when the training set size was fixed to 5% of the labeled pixels. In it all spectral bands were used and small classes were dismissed, that is, a 9-class problem was faced. In [1] the 16-class problem was tackled and a smaller number of bands was used but still 5% of the labeled dataset was needed to obtain an accuracy of 92%. Note that, in these works, when random pick is performed a priori probabilities of classes are kept and all classes are represented in the training set. Here the 16-class problem is faced with a very reduced training set. With the selection mode suggested in this paper, an accuracy of 96% can be obtained with only the 3.2% of the labeled pixels and

using only 3 spectral bands.



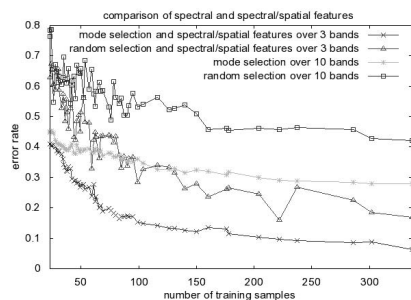
**Fig. 1.** Learning curves for different number of spectral features comparing the result selecting the training set with the corresponding number of training samples picked at random.



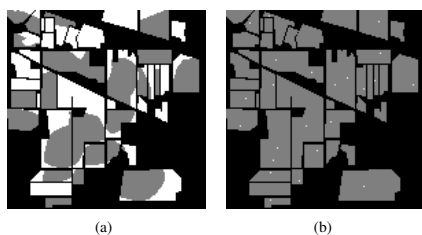
**Fig. 2.** Learning curves for different number of spectral bands using spectral/spatial features comparing the result selecting the training set with the corresponding number of training samples picked at random.

## VII. CONCLUSIONS

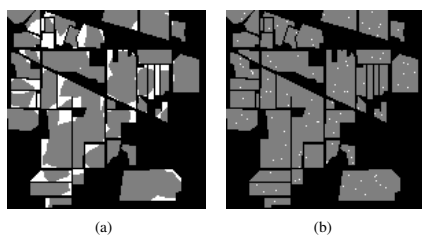
A semi-supervised segmentation and classification scheme has been suggested. Thanks to the mode selection performed by the clustering process, training samples are selected and only interesting samples are labeled by the expert. In this sense their collaboration is reduced while performance is raised in comparison with random selection and classification. Using a clustering method makes the result suitable for classifying with a simple nearest neighbor rule obtaining fairly goods results when fewer initial information is provided. Neither the process is computational expensive since it has been shown that not all spectral bands and not a high number of features were needed in our experiments. On the other hand, small classes may be missed by the clustering



**Fig. 3.** Learning curves resulting from selecting the training set and the corresponding number of training samples picked at random for spectral and spectral/spatial features.



**Fig. 4.** Classification results using 24 spectral/spatial features derived from 3 bands and 23 selected training samples. (a) representation of misclassified pixels in white and (b) training samples shown in white. Error rate was 0.41.



**Fig. 5.** Classification results using 24 spectral/spatial features derived from 3 bands and 104 selected training samples. (a) representation of misclassified pixels in white and (b) training samples shown in white. Error rate was 0.147.

procedure and then dismissed in the classification step. To tackle this problem the clustering step should be improved and probably a post-processing technique could also be of interest.

## VIII. REFERENCES

- [1] Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla, "Filter banks for hyperspectral pixel classification of satellite images," in *CIARP 2009, Lecture Notes in Computer Science*, vol. 5856, pp. 1039–1046, Springer, 2009.
- [2] A.Plaza, P.Martínez, J.Plaza, and R.Pérez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 43, pp. 466–479, 2005.
- [3] M.Fauvel, J.A.Benediktsson, J.Chanussot, and J.R.Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 46, no. 10, pp. 3804–3814, 2008.
- [4] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pattern Recogn.*, vol. 43, no. 7, pp. 2367–2379, 2010.
- [5] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 47, no. 8, pp. 2973–2987, 2009.
- [6] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Systems, Man, and Cybernetics*, pp. –, 2010.
- [7] RO Duda and PE Hart, *Pattern classification*, John-Wiley and Sons, 2001.
- [8] Y Cheng, "Mean shift, mode seek, and clustering," *IEEE Transaction on Pattern Analysis and Machine*, 1995.
- [9] M. Petrou and P. García-Sevilla, *Image Processing: Dealing with Texture*, John-Wiley and Sons, 1 edition, 2006.
- [10] Adolfo Martínez-Usó, Filiberto Pla, and Pedro García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 45, pp. 4158–4171, 2007.
- [11] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, Hoboken, NJ: Wiley, 1 edition, 2003.
- [12] A.Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote sensing of environment*, vol. 113, pp. 110–122, 2009.

## Selection of samples for active labeling in semi-supervised hyperspectral pixel classification

Olga Rajadell<sup>a</sup> and Cuong V. Dinh<sup>b</sup> and R. P. W. Duin<sup>b</sup> and Pedro García-Sevilla<sup>a</sup>

<sup>a</sup>University Jaume I, INIT, Spain

<sup>b</sup>Delft University of Technology, Pattern Recognition Laboratory, Delft, The Netherlands

### ABSTRACT

One of the problems in semi-supervised land classification tasks lies in improving classification results without increasing the number of pixels to be labeled. This would be possible if, instead of increasing the amount of data we increased the reliability of the data. We suggest to replace the random selection by a unsupervised clustering based selection strategy in building the training data. We use a mode seeking clustering method to search for cluster representatives, which will be labeled and then used for training. Here an improvement to the result of the clustering algorithm is introduced by taking advantage of the spatial information in the image. The number of selected samples provided by the clustering can be reduced by using a spatial-density criterion to dismiss redundant training information. Two different alternatives are considered for a spatial criterion, one dismisses selected samples in the same neighbourhood and the other includes the pixel coordinates for giving the spatial information a larger weigh in the clustering. Both alternatives improve the classification-segmentation results. The classification scheme with training selection provides state-of-the-art pixel classification results using a smaller training set and suggests an alternative to random selection.

**Keywords:** Pixel classification, hyperspectral imaging, semi-supervised classification, mode seeking.

### 1. INTRODUCTION

Clustering techniques allow us to divide data in a feature space into groups of similar objects. A very large number of clustering techniques is available. However, most of them rely upon a prior knowledge on the data, such as the number of clusters and the shape of clusters in the feature space (often elliptical). When dealing with an arbitrarily structured feature space, only nonparametric methods are applicable since no model assumptions can be made.<sup>1</sup> The methods can be distinguished into hierarchical and density based procedures. The first ones either aggregates or divides the data set according to some agreed measure. The latter considers the probability density function of the feature space and search for local maxima. Based on the local structure of the feature space, a number of samples are associated to the maxima found.<sup>2</sup>

Specially when samples represent pixels from an image, clustering algorithms have successfully been applied to image segmentation in various fields and applications. However, our purpose here is to segment and classify hyperspectral satellite images. Fully unsupervised procedures often have insufficient accurate segmentation result. For such a reason, a hybrid scenario between supervised and unsupervised techniques is often used. Semi-supervised learning methods are applied where some of the data points have labels. This scenario happens when data collection and feature extraction is automated but the labeling is done interactively. This is expensive both in time and cost. In this case the fewer labeled data the system can work with the better.<sup>3</sup>

Choosing data to be labeled has been a concern solved by randomly picking samples to move from an unsupervised scenario towards a semi-supervised one. This is easy, computationally cheap but not reliable, since among all the samples to be chosen random behavior can return a good, middle or very bad choice. There is also the widely used supervised solution in which samples are randomly picked within each class so previous knowledge about all classes is needed.

Reviewing data analysis techniques, they have proved their usefulness in providing relevant data when no prior knowledge is available. We suggest to use a clustering analysis to find samples of interest, ask an expert for their labels and train a classifier, as presented by Rajadell et al.<sup>4</sup> In this paper a further improvement of the selection is studied by taking advantage of the fact that samples are pixels in the image. Consequently spatial criteria added in two different ways will help to make a better selection and provide a much lower error than a random selection and still outperform the ones presented

---

E-mail: orajadel,pgarcia@lsi.uji.es, v.c.dinh@tudelft.nl, r.duin@ieee.org

in previous work.<sup>4</sup> This spatial criterion can be used after the clustering algorithm or within. In the first case, discarding selected samples within the same neighbourhood to get a smaller set of samples that spatially represent the same region in the image. Another possibility that is studied is including the pixel coordinates as features of the samples forcing the spatial location of a sample to get higher weight in the distances calculated between samples

A review of the clustering technique used can be found in Section 2 and the selection scheme<sup>4</sup> is summarized in Section 3. The spatial improvements are introduced in Section 4 and their results are shown and analyzed in Section 5. Conclusions and discussion are given in Section 6.

## 2. MODE SEEK CLUSTERING

Mean shift represents a general non-parametric mode clustering procedure. In contrast to the classic K-means clustering approach,<sup>5</sup> there are no embedded assumptions on the shape of the distribution nor the number of modes/clusters. Interest in mean-shift clustering was revived in 1995 by Cheng,<sup>6</sup> and Comaniciu et al.<sup>7</sup> further popularized it. A feature space is a multidimensional space in which each parameter considered to represent a sample (feature) is a dimension and the sample can be mapped as a point in that d-dimensional space. The main idea behind mean shift is to treat the points in the d-dimensional feature space as an empirical probability density function where dense regions in the feature space correspond to the local maxima or modes of the underlying distribution. For each data point in the feature space, one performs a gradient ascent procedure on the local estimated density until convergence. The stationary points of this procedure represent the modes of the distribution. Furthermore, the data points associated with the same stationary point (mode) are considered members of the same cluster.

Given  $n$  data points  $x_i \in \mathbb{R}^d$ , the multivariate kernel density estimate using a radially symmetric kernel  $K(x)$ , is given by,

$$\hat{f}_K = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad (1)$$

where  $h$ , bandwidth parameter, defines the radius of kernel. The radially symmetric kernel is defined as,

$$K(x) = c_k k(\|x\|^2), \quad (2)$$

where  $c_k$  represents a normalization constant. Taking the gradient of the density estimator  $\hat{f}_K$  and some further algebraic manipulation yields,

$$\nabla \hat{f}(x) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}} \left[ \sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \right]}_{factor1} \underbrace{\left[ \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \right]}_{factor2}, \quad (3)$$

where  $g(x) = -k'(x)$  denotes the derivative of the selected kernel profile. The first factor is proportional to the density estimate at  $x$  (computed with the kernel  $G = c_g(\|x\|^2)$ ). The second factor, called the mean shift vector,  $\mathbf{m}$ , points toward the direction of maximum increase in density and is proportional to the density gradient estimate at point  $x$  obtained with kernel  $K$ . The mean shift procedure for a given point  $x_i$  is as follows:

1. Compute the mean shift vector  $m(x_i^t)$ .
2. Translate density estimation window:  $x_i^{t+1} = x_i^t + m(x_i^t)$ .
3. Iterate steps 1. and 2. until convergence,  $\nabla f(x_i) = 0$

The most computationally expensive component of the mean shift procedure corresponds to the identification of the neighbours of a point in space (as defined by the kernel and its bandwidth). This computation becomes unwieldily for high dimensional feature spaces. Proposed solutions to this problem include employing approximate nearest-neighbour hashing-based search<sup>8</sup> in which a parameter  $s$  must be decided. In this case the density for each point is calculated for distances calculated in a neighbourhood  $s$ . Consequently, the size of the parameter  $s$  is inverse to the number of clusters that will be found;  $s$  is used to calculate the distances and those distances for the density associate to each point, the smaller the neighbourhood is the bigger number of local maxima would be found and surrounding these maxima, the correspondent clusters.

### 3. CLASSIFICATION SCHEME

Comaniciu et al. states<sup>7</sup> that vision tasks can be improved if they are supported by more reliable information. Nowadays input databases used for segmentation and classification of hyperspectral satellite images are highly reliable in terms of spectral and spatial resolution. Therefore, we can consider our feature space representation of the data is reliable too. However, in segmentation and classification of this kind of images, training sets are often built by randomly picking a percentage, against the principle of providing more reliable information. Here the proposed semi-supervised pixel classification scheme presented Rajadell et al.<sup>4</sup> is explained. The scheme makes an unsupervised selection of the training samples based on the analysis of the feature space. This succeeded in improving the training set and proceeded as follows:

1. In order to reduce the number of spectral bands to be used, a set of spectral bands, given a desired number, is selected by using the band selection method WaLuMi.<sup>9</sup>
2. Mode Seek clustering procedure is applied over that reduced feature space. An improvement in the clustering process is included by adding the spatial coordinates of each pixel in the image as features. This provides a spatial component very suitable for clustering since it is based in distances between samples.
3. The modes (centers of the clusters) resulting of the previous step define the training set for the next step. The expert is involved in this point by providing the corresponding labels of the selected samples. Here the expert is simulated by checking the labels in the ground truth provided for only those samples.
4. A  $KNN$  classifier with  $k = 1$  is build with the training set defined above. Note that at this point the spatial coordinates are dismissed as features. Regarding the clustering step it is always performed over the spectral features provided by the band selection method used. However, once the modes are obtained, the features to be used for the classification step can be the same or changed. In this paper, we will change to a spectral/spatial feature space for the classification step.

As explained in the previous Section, the parameter  $s$  of the clustering algorithm can be tuned to obtained a higher or lower number of clusters, that is, the number of interesting points to be labeled. The increase of this parameter is inverse to the number of clusters found. As it will be seen, the number of modes has a direct impact on the performance since they stand for the size of the training. Still the results are better than the ones obtained using a random picked for whatever the size of the training set.

### 4. SPATIAL IMPROVEMENT OF THE SELECTION

The nature of the feature space is application dependent. As many disadvantages they may have, there are also possibilities to explore if we take advantage of their own characteristics. In the problem we try to solve, our samples are pixels in the space. Given an image there is a pair of coordinates for each pixel in addition to all the features (measures) given by the spectrometer. Besides, land cover classification task count with the advantage that spatially connected samples are likely to belong to the same class, that is they are close in terms of spatial coordinates. In fact, there is only one group of samples that do not fulfil this statement, the borders of the class areas. Based on this two facts, we suggest two different ways of incorporating the spatial information into this scheme to improve the result of the selection of samples.

The clustering algorithm searches for local density maxima where the density function has been calculated using the distances for each sample in its  $s$  neighbourhood using a dissimilarity measure as distance between pairs of samples. According to this, large connected areas in the image space that represent a class can be split in several clusters when  $s$  is

not big enough to include all samples in that area. Also, classes that are spread in different areas in the image space can be included or not in the same cluster depending on how similar their features are. Whereas in the first case a few samples well located within the space area of the class are desired, when areas are located in different place in the image we would like to keep them as different clusters so different centers would be provided. Tuning the parameter  $s$  is not enough for getting a balance between the two targets. Smaller  $s$  help to detect more areas but unique big areas would have many unnecessary samples. On the other hand, larger  $s$  would provided not more than the necessary samples for big unique areas but other areas would be missed instead. So there is not such a balance tuning  $s$  when the problem is multi-class with unbalanced classes that may also lie spread along the image. Next, we suggest two different alternatives.

#### 4.1 Spatial criterion

Fig. 1.a shows the situation in which more samples than needed are selected. The centers of the clusters may stay close to each other in the spatial domain. Therefore, it results to the redundancy in training information. In our paper, we use a spatial criterion that only selects the cluster center which has the highest density among the cluster centers spatially connected. This criterion helps to merge small clusters together as they are likely to represent the same class in the hyperspectral data. Fig. 1.b shows the result after applying the spatial criterion. Notice that the two points next each other inside the purple area in the center of the image (next the green one), or in the orange area below the light green, in the bottom left part of the image, are reduced to one. The same happens to some other areas. It is important to observe in Figure 1.c that all of those pairs of samples belong to areas whose features seem to be not close enough and they are split in several clusters, thanks to this process they are now represented by one center. Among them, the sample chosen is the one with highest density since it guarantees to represent a higher amount of samples.

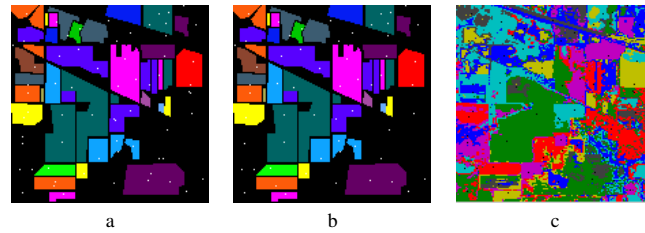


Figure 1. Clusters centers represented as white points over AVIRIS database groundtruth (a) 66 Cluster centers resulting from clustering procedure. (b) the previous result after applying the spatial criterion, 58 pixels remain. (c) the cluster result itself with the centers represented as black points (c).

#### 4.2 Clustering spatial improvement

Above we introduced the issue of redundant training information for points selected by the cluster algorithm as centers of clusters. It happened in spatial areas where the clustering was not able to find a homogeneous area and there was several small clusters that are very close in the image space. In this section, we aim at tackling this problem by embedding the spatial information directly into the clustering algorithm. Let us consider Eq.3. It is used for the search of the local maxima and the difference between samples is considered. In that difference, all features (dimensions) are considered and the spatial coordinates are two of those dimensions. For two samples that are close in terms of difference, if some feature is numerically enhanced to over-count in that calculation the difference in terms of that features is enhanced. We suggest to do such a thing with spatial coordinates. Multiplying coordinates by an arbitrary large number would make them more influent on the differences between samples, so when two samples were close spatially their distance is closer and the way round. Such a number should be decided in terms of the range of the features provided by the spectrometer so the coordinates are overweighed but they do not cause the rest of features be dismissed in the difference. In Figures 2.c-d the cluster result is represented for the case of normal mode seek cluster (c) and the case in which high weights are assigned to coordinate features (d). The main difference is the homogeneity of the clusters in the space as a direct consequence of giving a higher importance to coordinates. Notice that bottom right side of the image in Figure 2.c is messy, several clusters are involved in the same area. In Figure 2.d that area is covered by three clear clusters instead. Therefore, the features

were not clear enough to split it into spatial areas and thanks to introducing the overweigh spatial coordinates they are now spatially separated. From the image segmentation point of view, this may not be a proper strategy since the noisy areas obtained may be due to the heterogenous nature of this part of the image. However, our aim is to get nicely distributed centers and we know that connected areas are likely to belong to the same class. Besides, this can avoid our training set to have redundant information, though. Consequently, noticeable differences can be found between Figures 2.a-b: centers are more distributed and areas that were missed are now found (look blue area on the center top of the image). Because the selection using only a pre-clustering strategy proved to work better than random, it is expected that alternatives enhancing the role of coordinates also improve random and hopefully normal cluster scheme version.

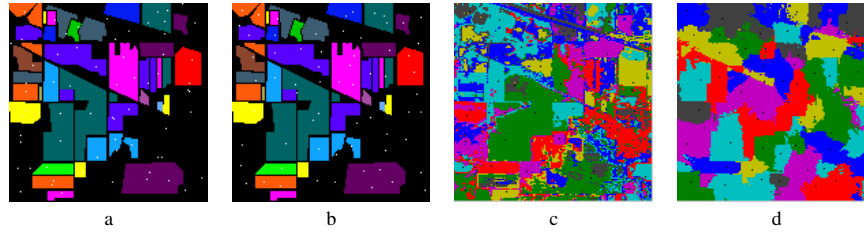


Figure 2. Cluster centers represented as white points over AVIRIS database groundtruth (a) 66 Cluster centers resulting from clustering procedure. (b) 66 centers obtained using a different  $s$  with clustering overweighed coordinates. Cluster result with the centers represented as black points, (c) for the first case and (d) for the second.

## 5. RESULTS

### 5.1 Databases

A widely used database has been used in the experiments (see Fig. 3). Hyper-spectral image data 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR).<sup>10</sup> The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. In it, three different growing states of soya can be found, together with other three different growing states of corn. Woods, pasture and trees are the bigger classes in terms of number of samples (pixels). Smaller classes can be also found such as steel towers, hay-windrowed, alfafa, drives, oats, grass and wheat. In total, AVIRIS has 16 clases labeled but part of the image is unlabeled which is known as the background. This so called background will be here considered as an heterogenous class.



Figure 3. AVIRIS database with its groundtruth.

### 5.2 Experimental setup

The dataset was reduced to 10 bands using the band method selection named in Section 3. The parameter  $s$  in the mode seeking algorithm is varied to get a learning curve. Remember that the smaller  $s$  is the bigger the training set and vice versa. We use the KNN with  $k=1$  as the classifier. We note that it is not an arbitrary choice. Taking into account that



the clustering procedure is based on density estimation on a dissimilarity space, the local maxima correspond to samples which minimize its dissimilarity and has a high amount of samples close to it in terms of distance. So these samples selected are highly representative in classifiers that calculate distances. On the other hand, for those that search frontiers, this method will not provide useful training data. Notice that for the random pick strategy these figures are the mean of 10 classification attempts to increase the stability of the random results. Moving to the parameters related to the spatial criteria suggested here, the neighbourhood in which a center should be dismissed depends on the size of the image and the size of the classes. A big image with big class areas would need a bigger neighbourhood than a smaller image since the neighbourhood used before may skip class areas in between for the smaller one. A neighbourhood of  $9 \times 9$  was chosen. As for the number to enhance the coordinates, since the features of our database was ranged  $[0..255]$  several alternatives were tested and  $me[5..20]$  was found to provide the same results so the factor to enhance them was set to 10. Regarding the feature space for classification, this is changed after clustering. Coordinates are dismissed and spatial/spectral features suggested by Rajadell et al.<sup>11</sup> are used. To calculate these features, 3 bands, 4 orientations and 2 frequency scales are used.

### 5.3 Results in figures

Results are presented in Fig. 4.a when  $s$  parameter is progressively decreased and as consequence the training data increases. In all cases more training data increase the performance of the classifier but not all methods of including data provide the same learning gain. The distance between random selection and the rest of alternatives is noticeable. It proves that an analysis of the data is preferable to a blind random pick because there is a selection to provide suitable representant as training.

The initial clustering based scheme without any improvement, although it is better than random selection can still be improved since redundant training data may be included. By either discarding centers in a given neighbourhood or enhancing the role of spatial coordinates the results outperforms the original scheme. Comparing the two alternatives, the second one gets a better improvement. It is interesting to point out that it is worthless to try to discard centers in a neighbourhood when coordinates has already been enhanced. Unless the neighbourhood chosen in the clustering algorithm is so small that provides loads of centers and the neighbourhood for the criteria is large in comparison, the clustering will already provide centers spatially far. That is the reason why plots "coordinates overweighed + discarding neighbourhood" and "coordinates overweighed" provide the same results.

Observe that at the beginning no differences are found between the alternative improvements and the normal scheme version. This is due to the fact that few samples selected mean few clusters as a result of the clustering and in that situation centers are rarely place nearby. As the size of the training data increases, differences are found. This can also be observed in Fig. 4.b. In Fig. 4.b the effect of the parameter  $s$  is studied. As was previously said, smaller  $s$  provides a higher amount of local maxima and consequently more clusters and more centers. Notice also that the normal strategy and the one that discard selected samples in a given neighbourhood stay together until the number of centers grows and some neighbouring centers appear. However when the improve is included in the clustering data by overweighing the coordinates, the result of the clustering provides a high number of clusters (centers) since enhancing the role of the spatial coordinates force the clusters to split when samples are spatially away. Last, as stated before, it is worthless to include a post-discarding process since the clusters are already forced to be spatially away and that's why the number of centers of the option "overweighed coordinates + discarding neighbourhood" (represented with points) appears on top of the improvement overweighed coordinates.

### 5.4 Segmentation results

The improvements in error rate are interesting but do not give an overview of what happens in the image in terms of class recognition and segmentation. Here we analyze the results in the space domain. First, observe in Figure 5 a case with a reduce number of training samples (71). In Figure 5.a, selected pixels come directly from the result of the clustering using  $s = 56$ , in Figure 5.b they result from the clustering using  $s = 44$  and performing a neighbourhood discarding, as a consequence the same number of selected pixels as before remain and in Figure 5.c the selected pixels are found using clustering with  $s = 91$  but in this case coordinates were overweighed. For getting the same amount of selected samples one should use a bigger  $s$  when discarding the neighbourhood to force the clustering to provide more clusters centers and then discard the redundant ones, as for the case in which coordinates are overweighed enhancing the role of the coordinates in the distance calculation make clusters split when samples are spatially away and that is why to get the same number of clusters as the other two alternatives, a greater  $s$  is needed (remember that the bigger  $s$  the smaller number of clusters).

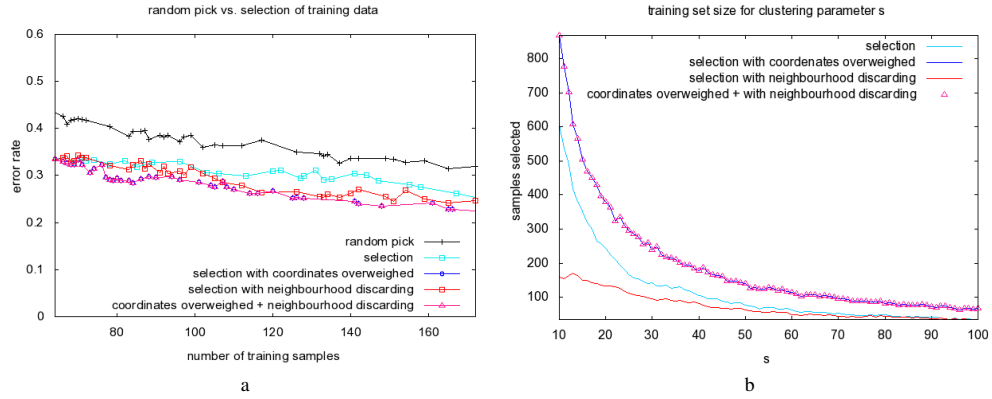


Figure 4. (a) Learning curve of the KNN classifier in terms of error rate when increasing the size of the training data in number of samples selected by the scheme suggested with the two improvement alternatives compared with the usual random pick. (b) Effect of the parameter  $s$  on the size of the number of samples selected. Both using AVIRIS database.

To see how this is translated into classifications results, look the second row of Figure 5. The corresponding results can be seen where misclassified pixels are represented in white. Notice that only when coordinates are overweighed the blue area in the top center of the image is selected and included for training whereas the rest keeps more or less the same.

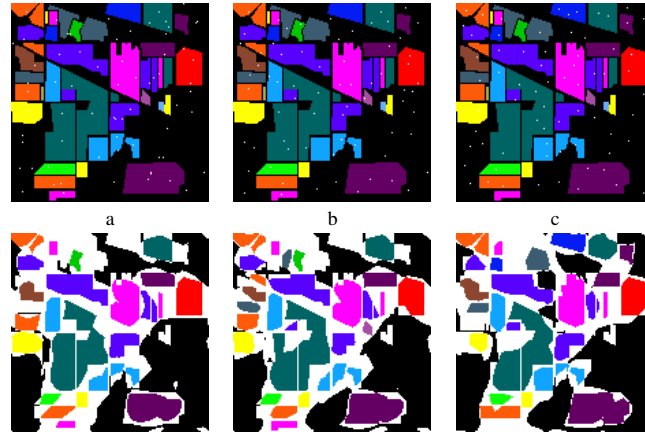


Figure 5. First row: 70 Pixels selected for training by (a) simply clustering, (b) clustering and discarding those lying in the same neighbourhood and (c) clustering overweighing the coordinates of each sample. Second row: Corresponding classification results for each method respectively.

The database has 21025 samples, that is, 70 samples represent the 0.33% of the data. We should move to a higher percentage. Let's consider the 2% and the best of our improvements here. See results in Figure 6.a and compare with random pick results in Figure 6.b. Whereas one random performance obtains a error rate of 0.23351 selecting samples for training reaches an error rate of 0.1763. Keep in mind that here the black area has also been considered as a class, so this

is a 17-class problem and the whole image has been classified. Observe the difference in the left top part of the image where random selection missed most of the classes and the selection performs recognizing all of them. The classification-segmentation result itself can be seen in Figure 7.



Figure 6. Error representation in white for a classification task using 2% of the data for training (a) selecting samples by clustering using coordinates overweighted. (b) by randomly picking samples for training.



Figure 7. Segmentation-classification result using the selection for building the training set with overweighted coordinates.

## 6. CONCLUSIONS

A spatial improvement to the general pixel classification scheme previously presented<sup>4</sup> has been suggested. There, a replacement to the random pick was suggested in order to build a training set based on an unsupervised study of the feature space. In this paper, the training set selection is improved by avoiding the redundancy in the training data. Two possibilities were introduced, first neglecting redundant information included in the result of the clustering and second enhancing the coordinates in the calculation of the differences to avoid small clusters. Both alternatives improved the initial selection. Because the first alternative is performed over the original as a post-process, the original and this first improvement are equal until the number of clusters increase significantly and redundant information is included. Whereas, the second one provides a different clustering result and performs better than the initial and the first alternative from the beginning. Regarding the target of the initial suggestion, selecting training data is convenient, devices make feature space reliable and an analysis of the data can provide a better starting point than random selection for the classification task.

## ACKNOWLEDGMENTS

Thanks to Fundació Caixa Castelló-Bancaixa for funding by grant FPI PREDOC/2007/20. Also to the Spanish Ministry of Science and Innovation for supporting in projects CSD2007-00018 (Consolider Ingenio 2010), AYA2008-05965-C04-04 and MTM2009-14500-C02-02.

## REFERENCES

- [1] Jain, A., Duin, R., and Mao, J., "Statistical pattern recognition: a review," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22**, 4–37 (jan 2000).

- [2] Wilson, R. and Spann, M., "A new approach to clustering," *Pattern Recognition* **23**(12), 1413–1425 (1990).
- [3] Ng, A., Jordan, M., and Weiss, Y., "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems* **14**(5), 849–856 (2002).
- [4] Rajadell, O., Dinh, V., Duin, R. P. W., and García-Sevilla, P., "Semi-supervised hyperspectral pixel classification using interactive labeling," in [*Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2011], (june 2011).
- [5] Duda, R. and Hart, P., [*Pattern classification*], John-Wiley and Sons (2001).
- [6] Cheng, Y., "Mean shift, mode seek, and clustering," *IEEE Transaction on Pattern Analysis and Machine* **17**, 790–799 (Aug. 1995).
- [7] Comaniciu, D. and Meer, P., "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**, 603–619 (may 2002).
- [8] Georgescu, B., Shimshoni, I., and Meer, P., "Mean shift based clustering in high dimensions: a texture classification example," *Proceedings. Ninth IEEE International Conference on Computer Vision* **1**, 456–463 (oct 2003).
- [9] Martínez-Usó, A., Pla, F., and García-Sevilla, P., "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. on Geoscience & Remote Sensing* **45**, 4158–4171 (2007).
- [10] Landgrebe, D. A., [*Signal Theory Methods in Multispectral Remote Sensing*], Hoboken, NJ: Wiley, 1 ed. (2003).
- [11] Rajadell, O., García-Sevilla, P., and Pla, F., "Filter banks for hyperspectral pixel classification of satellite images," in [*CIARP 2009, Lecture Notes in Computer Science*], **5856**, 1039–1046, Springer (2009).

# Improving hyperspectral pixel classification with unsupervised training data selection

Olga Rajadell, Pedro García-Sevilla, Viet Cuong Dinh and Robert P.W. Duin

**Abstract**—An unsupervised method for selecting training data is suggested here. The method is tested by applying it to hyperspectral land-use classification. The data set is reduced using an unsupervised band selection method and then clustered with a non parametric cluster technique. The cluster technique provides centers of the clusters and those are the samples selected to compose the training set. Both the band selection and the clustering are unsupervised techniques. Afterwards an expert labels those samples and the rest of unlabeled data can be classified. The inclusion of the selection step, although unsupervised, allows to select automatically the most suitable pixels to build the classifier. This reduces the expert effort because less pixels need to be labeled. However, the classification results are significantly improved in comparison with results obtained by a random selection of training samples, in particular for very small training sets.

## I. INTRODUCTION

Segmentation and classification are well known issues in image processing that are lately faced as a single problem by using pixel classification. For classification, expert labeling is needed to train the system to later classify unlabeled samples. Some authors work in a supervised scenario where prior knowledge is available and training data is selected within each class [1] [2]. Active learning techniques have also been applied. In these, the expert collaboration improves progressively the training data [3] [4]. In both cases, the way the training data is first selected is a concern generally solved by randomly picking among the unlabeled data. This is unsupervised but not very efficient. Randomly distributed samples can lie in non interesting areas and reducing the size of the training set may make the training data non representative. On top of that, expert collaboration is expensive. To face both problems we suggest to provide the system with the most interesting samples from the beginning. The traditional randomly selected training set is thereby replaced by a selective choice.

In unsupervised scenarios, data analysis techniques are widely used for finding relevant data when no prior knowledge is available. Among them, clustering techniques allow to divide data into groups of similar samples. A very large number of clustering techniques is available but some of them rely upon a prior knowledge, such as the number of clusters and the shape of clusters in the feature space (often elliptical). When dealing with an arbitrarily structured feature space, only nonparametric methods are applicable since no model

assumption have to be made [5]. Clustering algorithms have successfully been applied to image segmentation in various fields and applications [6]. Fully unsupervised procedures often have insufficiently accurate segmentation results. For such a reason, a hybrid scenario between supervised and unsupervised techniques is of high interest. In this case, the methods applied use a small set of labels to train a classifier. Because labeling is neither fast nor cheap, the fewer labeled data the system needs the better [7].

The contribution of this paper is the introduction of a method to select the training data. The suggested method is tested for hyperspectral landscape image classification and compared with a random selection of the training set. Results based on a selective choice of the training set outperform those achieved with randomly picked training data, mainly when a very small number of labeled samples is used. The scheme is presented in Section II with a focus on the selection method. Results will be shown over the dataset presented in Section III and analyzed in Section IV. Section V are conclusions.

## II. CLASSIFICATION SCHEME

Comaniciu et al. states in [8] that vision tasks can be improved if they are supported by more reliable data. Nowadays databases used for segmentation and classification of hyperspectral satellite images are fairly reliable in terms of spectral and spatial resolution. Therefore, we can consider that our feature space representation of the data is reliable. However, training sets are often built by randomly picking a percentage of samples. We suggest to make an unsupervised selection of the training samples based on the analysis of the feature space. This aims at providing an improved training set. The whole classification scheme proceeds as follows:

- 1) A band selection method is used. With it the data set is reduced to a smaller set of bands. This set is less correlated than the original while it keeps as much information as possible. We used the WALUMI band selection method [9], but any other band selection method that fulfils that requirement could be used instead.
- 2) A clustering procedure is applied over the reduced dataset. The centers of the clusters found form the selected training set. A non-parametric clustering technique is used and prior knowledge is not needed.
- 3) The expert is involved once, after the selection, to provide the corresponding labels of the selected samples. In this paper the expert will be simulated by checking the corresponding labels on the groundtruth.
- 4) A classifier is built using the training set defined before. Although the clustering is performed using spectral

Olga Rajadell and Pedro García-Sevilla are with the University Jaume I, Spain, within the Institute of New Imaging Technologies (<http://www.init.uji.es>). Viet Cuong Dinh and Robert P.W. Duin are with Delft University of Technology, PRLab, The Netherlands. Viet Cuong Dinh is also with of CTR, Villach, Austria.

features, we test that the selection obtained can be used independently to the type of features used for classifying.

#### A. Mode seeking clustering

Mode seeking clustering is a well known clustering principle for image segmentation. Based on a given set of objects, in case of images these are the pixels, a non-parametric estimate of the probability density function (pdf) is made. The modes of this pdf correspond to the clusters. In a gradient search all objects are used as a starting point and objects ending up in the same mode belong to the same cluster. Neither the number of clusters nor their shape has to be predefined.

The most popular mode seeking procedure is the mean shift algorithm [10] [11]. It is based on a Parzen kernel density estimate of the pdf. In contrast to the classic K-means clustering [12], or the more advanced Mixture-Of-Gaussian density estimates there are no embedded assumptions on an underlying Gaussian distribution of the data [10] [8]. In the mean shift algorithm the direction of the local gradient is found by a shift of the mean of the local mean when the distances to the objects in a local neighborhood are weighted by the chosen kernel. This procedure works well for the segmentation of color images, especially when some spatial information is included in features representing the pixels [8]. Problems with mean shift are that the modes as well as the convergence are not sharply defined. Thereby, separate nearby modes may be found that are erroneously not merged. Moreover, formally all pixels have to be used as a starting point, which is very time consuming.

Another algorithm based on mode seeking is  $k$ NN mode seeking. Instead of the Parzen kernel density estimate it is entirely based on the distances to the  $k$ -th neighbor. It can be traced back to a proposal by Koontz et al. in 1977 [13]. It has been around in the Matlab toolbox PRTools [14] for 20 years. Recently it has been redefined [15] and compared with mean shift. The procedure can be summarized as:

Do for all objects:

- 1) Find its  $k$  nearest neighbors.
- 2) Use the distance to the  $k$ -th neighbor as a measure for the density (in fact one over the distance).
- 3) Define a pointer to the object with the highest density in the  $k$ -neighborhood.
- 4) Follow from all objects the pointers until objects are reached that point to themselves: the modes.

Various implementations are studied. We used one that is based on an approximate nearest neighbor search [16]. It performs the above algorithm for clustering 10366 objects in 5 dimensions with  $k=100$  in 1.4 seconds and with  $k=10$  in less than a second (0.7) on a standard PC (Intel Core Duo 2GHz, with 4GB of RAM). Its computational complexity is about  $O(kn^2)$  for data sets with  $n$  objects. The dependency on the dimensionality is heavily problem dependent due to the approximate nearest neighbor. Advantages of this algorithm over mean shift are that it is much faster and converges exactly to modes that correspond with objects (pixels). Moreover it can handle high dimensional spaces and finds solutions for sets of

$k$ -values in almost the same time as needed for the largest  $k$ -value in the set.

#### B. The role of spatial coordinates

The specific task targeted here is the classification of land cover images. In this type of images, the samples are pixels and the classes the different areas in the image. Thus, samples within the same class are spatially connected (class connection principle or smoothness). This is an advantage because it adds extra information to the spectral information provided by sensors. However, it can happen that a class is located in more than one spatial location. In such a case, even being the same class, the characteristics of their samples can differ due to different lighting or soil conditions in the different locations.

The clustering algorithm chosen searches for local density maxima where the density function has been calculated using the distances for each sample in its  $k$  neighbourhood. A smaller  $k$  results in a higher number of clusters, that is helpful if we aim to select more samples from different areas. However, unique large areas would also have many samples selected within the same region that are unnecessary (redundant training data). On the contrary, bigger  $k$  would provide fewer selected samples for big areas but smaller areas or different locations of the same class would be missed instead.

We suggest to incorporate spatial information to the selection algorithm. Like this the clustering will also take into account their spatial connectivity. This has already been done in literature [17] by simply adding the spatial coordinates to the feature vector of each pixel. By adding the coordinates within the distance computation, samples nearby will have a higher probability of being clustered together and the opposite for spatially remote samples even if they belong to the same class. Note that coordinates are only used for this clustering step and only the spectral information (without the coordinates) or features derived from the spectral information are used in the classification step. This allows a fair comparison with several methods proposed by other authors and the random selection method included in the paper. That is, the features used in the classification step for the mode selection method and for the random selection method are exactly the same. The only difference lies in how the training set is built.

#### C. Spectral-spatial features

The contribution of this paper is a training selection method. Such a method should point out which samples are significant for training independently of the features used for classification afterwards. To that end, we suggest to switch the features for classification, using the same selected samples for training in order to show that this selection still outperforms a random pick selection. We choose a different type of features, spatial features extracted by filtering suggested in [18]. These are obtained by filtering the input image with a set of two-dimensional Gabor filters. The outputs of each pixel in the image forms its feature vector. Each Gabor filter is characterized by a preferred orientation and a preferred spatial frequency (scale) so this features characterized the texture contained in the image.

### III. DATASET

The dataset used in the experiments is widely known in the field. Hyper-spectral image 92AV3C (Fig. 1) was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR) [19]. The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. Classes range from 20 to 2468 pixels. In it, three different growing states of soya can be found, together with other three different growing states of corn. Woods, pasture and trees are the bigger classes in terms of number of samples (pixels). Smaller classes are steel towers, hay-windrowed, alfalfa, drives, oats, grass and wheat. In total, the dataset has 16 labeled classes.



Fig. 1. AVIRIS database color composition and groundtruth.

### IV. RESULTS

For all experiments, clustering is carried out using different values of the parameter  $k$  to get different sizes of training sets (selected samples). Notice that this is not an iterative process. The clustering is performed once and, as a consequence of the value of the parameter  $k$ , a number of samples is selected. The expert labels these samples and the classification is performed using only that labeled data as training and the rest as test. Plots in Fig. 2, 3, and 4 are represented in terms of error rate versus number of labeled samples provided for training. They represent the improvement of the classification when increasing the amount of labeled data.

A K-NN with  $K = 1$  classifier has been used (not to be confused with the  $k$ -NN mode seeking procedure used for clustering). This is not an arbitrary choice. Because the clustering procedure used is based on densities determined by distances, the local maxima (the pixels used for training) correspond to samples which have many objects in their direct neighborhood. Small classes, or uni-modal classes may be represented by a single training point, so larger values of  $K$  are not possible.

The dataset was reduced to different number of selected bands using WaLuMi band selection method. The bands selected used for the experiments carried out are presented in Table I.

#### A. Classification results

In Fig. 2 the learning curves for a different number of spectral bands are presented in both cases, selecting samples

no. of bands	selected bands
3	4, 67, 87
10	4, 24, 51, 67, 78, 87, 99, 118, 129, 182
20	4, 15, 24, 33, 35, 36, 41, 51, 67, 77, 79, 87, 95, 99, 111, 118, 129, 172, 182, 204

TABLE I  
SELECTED BANDS USING WaLuMi FOR AVIRIS DATASET.

with the method and picking the same amount of samples at random. It is noticeable that in all cases, when selecting the training set, the classification rate outperforms the result obtain when the same amount is picked at random. When a small training set is used the difference between the the error rate selecting and not selecting is 0.3, whereas it decreases to 0.15 when the training set grows. This happens because the higher the number of samples is picked, the chances of randomly select samples from all classes are bigger. Also when the number of samples to select is very small random is very unstable. Note that no advantage is obtained in involving a higher number of spectral bands in the process. However, the difference between using 10 and 20 bands is an increase of 10 features in the feature vector. The main reason for selecting information is that, once known which are the most informative bands for a given sensor, further repetitions of the same task can be performed dismissing information that was proved to be redundant for that task using that sensor.

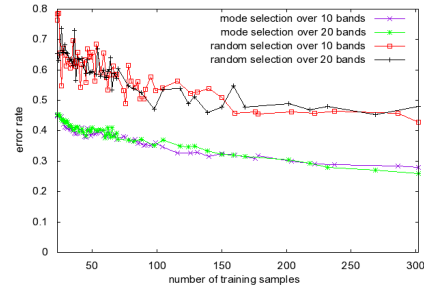


Fig. 2. Learning curves for different number of spectral features comparing the result selecting the training set with the corresponding number of training samples picked at random.

For the next experiment, the spatial-textural type of features is also used for classification. Note that the selection is the same and the target is to validate that the same training selection result improves the random selection being independently of the features used. These other features are computed from each band independently and 8 features are obtained per band. In Fig. 3 we show the learning curves obtained for the experiments that use 3 and 10 bands. Despite the difference between the size of the feature vector (24 for 3 bands and 80 for 10), no performance increase is noticed. As a summary also the difference in the error rate caused by changing the features for classifying (10 spectral and 24 spectral/spatial features) can be observed in Figure 4. Note how in both cases the error rate obtained using the random selection stays above the classifica-

tion including the training selection method. It is remarkable that both sets of features start around the same error rate but the difference is quickly introduced when more samples are included. When using spectral/spatial features the error rate decreases considerably. The characterization improvement that these features introduce, together with providing representative labeled data, obtains a fairly good well classified area with a relatively small amount of labeled data.

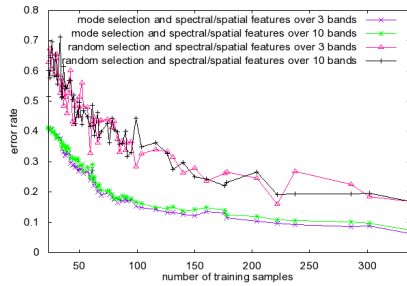


Fig. 3. Learning curves for different number of spectral bands using spectral/spatial features. Results selecting the training set are compared with the same amount of samples picked at random.

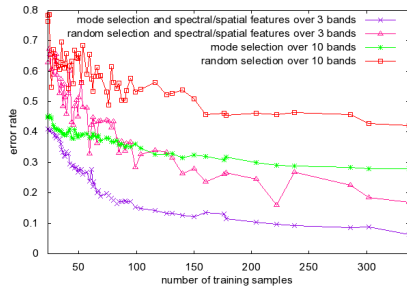


Fig. 4. Comparison between two types of features. Learning curves for the classification results using selection of the training and random pick.

### B. Analysis per class

Note that because classes are highly unbalanced, an increase in the performance is wanted when it represents an improvement for all classes and, in this case, large classes have a higher impact on the overall accuracy. In Table II the error rate per class is shown. The results obtained with 3.5% of labeled samples are comparable, in terms of per class accuracy, with results obtained in other scenarios using 10% of random labeled samples for training [1] or a fixed number of labeled samples per class (50 samples per class, 15 for small ones) [20]. This last approach favors small classes in comparison with the unsupervised selection method presented here. The number of samples per class used in the training set is here unsupervised and no prior knowledge is used.

Despite this disadvantage, the accuracy for very small classes is better than experiments where the training selection is not used. Stone-steel towers, alfalfa, grass/pasture-mowed have error rates around 0.10 with only one or two samples for training. Other classes usually dismissed in the classification experiments because of their size [2] [21] like wheat, corn and Bldg-Grass-Tree-Drives have error rates of 0.07, 0.14 and 0.01 using only six, nine and ten labeled samples.

classes	0.6% of training data		3.5% of training data	
	training/total	error	training/total	error
Stone-steel towers	1/95	0.04	2/95	0.05
Hay-windrowed	4/489	0.03	19/489	0.03
Corn-min till	6/834	0.33	27/834	0.17
Soybeans-no till	7/968	0.10	29/968	0.11
Alfalfa	1/54	0.07	2/54	0.11
Soybeans-clean till	4/614	0.40	21/614	0.12
Grass/pasture	4/497	0.14	14/497	0.21
Woods	9/1294	0.08	47/1294	0.04
Bldg-Grass-Tree-Drives	4/380	0.002	10/380	0.01
Grass/pasture-mowed	0/26	1	1/26	0.04
Corn	1/234	0.38	9/234	0.14
Oats	0/20	1	0/20	1
Corn-no till	8/1434	0.25	44/1434	0.13
Soybeans-min till	11/2468	0.21	90/2468	0.04
Grass/trees	5/747	0.11	28/747	0.06
Wheat	2/212	0.15	6/212	0.07
Overall error		0.26		0.12

TABLE II  
ACCURACY PER CLASS FOR THE 16 CLASSES CLASSIFICATION OF THE AVIRIS DATASET SELECTING THE TRAINING SET OVER THE SPECTRAL FEATURES CONCATENATED WITH THE SPATIAL COORDINATES AND CLASSIFYING USING SPATIAL-SPECTRAL FEATURES.

For an overview of the per class result, observe in Fig. 5.(a)(c) the selected training set (white points represented on the groundtruth) and the corresponding per class results Fig. 5.(b)(d)(where the color areas are well-classified pixels and the white ones miss-classified pixels). Both cases result from selecting training data by clustering over 10 spectral features plus two spatial coordinates, label the samples selected and use them as training set for a KNN classifier, replacing the spectral features by 24 spatial-spectral features for classification.

The case of a reduced number of training samples, Fig. 5(first row), demonstrates that one sample is needed to recognize a class (those areas where a mode is not found are dismissed in the classification). Leaving aside the misclassified areas, observe that those where a sample is labeled provide a well-classified region around them with 23 samples of training (only a 0.22% of the dataset). In a scenario with a very reduced amount of labeled data, it should be considered the possibility that the expert corrects the loss by adding training samples of a region that has not been detected. For the bigger training set size, Fig. 5(second row) with only 104 samples, fulfils that classes distributed in different regions have samples for each of those regions and big regions have several labeled samples distributed along the fields, maximizing the classified area.

### V. CONCLUSIONS

A method for selecting the training set has been suggested to replace the common random pick selection. This is useful when no prior knowledge is available and expert collaboration



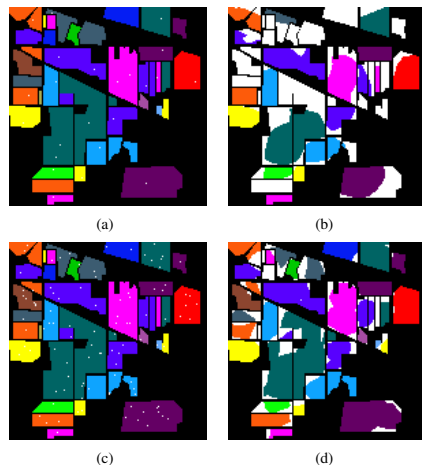


Fig. 5. Classification results using 24 spectral/spatial features derived from 3 bands. (a) 23 selected training samples shown in white. (b) representation of misclassified pixels in white, error rate was 0.41. (c) 104 selected training samples shown in white. (d) representation of misclassified pixels in white, error rate of 0.147.

is limited. Thanks to the selection of the training set, only relevant samples can be shown to the expert to be labeled. In this sense, expert collaboration is reduced while performance has shown to be raised in comparison with random selection. The method is based on an unsupervised study of the data by a clustering technique. Besides, a spatial improvement was suggested to avoid redundant training data by including spatial coordinates in the clustering process. This forced clusters to merge or split according to the class connection principle. Thus, the training set is representative and free of redundancies. The selection has shown to be valid for building a classifier even if the features are changed. It was shown that textural-spatial features can also benefit from this selection scheme and achieve same results with less training data. Indeed, results shown outperform results of classification methods in literature that use a random selection of their training set. On the top of that, the process does not need large amounts of data since it has been shown that not all spectral bands and not a high number of features were needed in our experiments.

#### ACKNOWLEDGMENT

Thanks to Fundació Caixa Castelló-Bancaixa for funding by grant FPI PREDOC/2007/20. Also to the Spanish Ministry of Science and Innovation for supporting in projects CSD2007-00018 (Consolider Ingenio 2010), AYA2008-05965-C04-04 and MTM2009-14500-C02-02.

#### REFERENCES

- [1] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pat.Recogn.*, vol. 43, no. 7, pp. 2367–2379, 2010.

- [2] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote sensing of environment*, vol. 113, pp. 110–122, 2009.
- [3] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 –2232, July 2009.
- [4] J. Li, J. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085 –4098, Nov. 2010.
- [5] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4 –37, Jan. 2000.
- [6] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 176 – 190, 2008.
- [7] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, no. 5, pp. 849 –856, 2002.
- [8] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 –619, May 2002.
- [9] A. Martínez-Usó, F. Pla, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 45, pp. 4158–4171, 2007.
- [10] Y. Cheng, "Mean shift, mode seek, and clustering," *IEEE Transaction on Pattern Analysis and Machine*, vol. 17, no. 8, pp. 790 –799, Aug. 1995.
- [11] K. Fukunaga and L. Hostettler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1977.
- [12] R. Duda and P. Hart, *Pattern classification*. John-Wiley and Sons, 2001.
- [13] W. Koontz, P. Narendra, and K. Fukunaga, "A graph-theoretic approach to nonparametric cluster analysis," *IEEE Transactions on Computer*, vol. 25, pp. 936–944, 1976.
- [14] R. Duin, D. de Ridder, P. Juszczak, C. Lai, P. Paclik, E. Pekalska, and D. Tax, "Ptools4," 2010. [Online]. Available: <http://prtools.org>
- [15] R. Duin, A. Fred, M. Loog, and E. Pekalska, "Mode seeking clustering by knn and mean shift evaluated," in *SSPR SPR 2012 Lecture Notes in Computer Science*, vol. 7626. Springer, 2012, pp. 51–59.
- [16] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *Journal of the ACM*, vol. 45, no. 6, pp. 891–923, 1998.
- [17] N. Pal and S. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, pp. 1277 – 1294, 1993.
- [18] O. Rajadell, P. García-Sevilla, and F. Pla, "Spectral-spatial pixel characterization using gabor filters for hyperspectral image classification," *IEEE Geoscience & Remote Sensing Letters*, vol. 10, no. 4, 2013.
- [19] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, 1st ed. Hoboken, NJ: Wiley, 2003.
- [20] Y.Tarabalka, J.Chanussot, and J.A.Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 40, no. 5, pp. 1267 –1279, Oct. 2010.
- [21] G.Camps-Valls and L.Bruzzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. on Geoscience & Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

## Training Selection with Label Propagation for Semi-supervised Land Classification and Segmentation of Satellite Images

Olga Rajadell and Pedro García-Sevilla\*

Institute of New Imaging Technologies, University Jaume I, Spain  
{orajadel,pgarcia}@lsi.uji.es

**Abstract.** Different scenarios can be found in land classification and segmentation of satellite images. First, when prior knowledge is available, the training data is generally selected by randomly picking samples within classes. When no prior knowledge is available the system can pick samples at random among all unlabeled data, which is highly unreliable or it can rely on the expert collaboration to improve progressively the training data applying an active learning function. We suggest a scheme to tackle the lack of prior knowledge without actively involving the expert, whose collaboration may be expensive. The proposed scheme uses a clustering technique to analyze the feature space and find the most representative samples for being labeled. In this case the expert is just involved in labeling once a reliable training data set for being representative of the feature space. Once the training set is labeled by the expert, different classifiers may be built to process the rest of samples. Three different approaches are presented in this paper: the result of the clustering process, a distance based classifier, and support vector machines (SVM).

**Keywords:** Semi-supervised classification, Image segmentation, Hyper-spectral imaging, mode seek clustering.

### 1 Introduction

The classification and segmentation of land usage in satellite images generally requires an expert who provides the corresponding labels for the different areas in the images. Some authors work with prior knowledge in a supervised scenario and training data is selected within each class [1][2]. Lately the research interest in active learning techniques, which move to a semi-supervised scenario, is raising. In new real databases, the expert labeling involves whether prior knowledge or checking at the land place itself, which could be highly expensive. The expert collaboration may be needed an unknown number of steps to improve the classification by helping in the training selection until the convergence condition is achieved [3][4]. Hence, the expert collaboration can be highly expensive and picking at random among the unlabeled pool is not convenient

---

\* This work has been partly supported by grant FPI PREDOC/2007/20 from Fundació Caixa Castelló-Bancaixa and projects CSD2007-00018 (Consolider Ingenio 2010) and AYA2008-05965-C04-04 from the Spanish Ministry of Science and Innovation.

182 O. Rajadell and P. García-Sevilla

because classes are often very unbalanced and the probabilities of getting an efficient representative training data is inverse to the amount of labeled samples. Consequently, decreasing the size of labeled data is a problem. Whereas for classifier based on distances, larger training sets overfit our classifier and it is preferable to provide the classifier with a few interesting highly descriptive samples [5]; for other types of classifiers providing a considerable amount of training samples is a concern.

In unsupervised scenarios, data analysis techniques have proved being good at providing relevant data when no prior knowledge is available. Among them, clustering techniques allow us to divide data in groups of similar samples. Specially when samples represent pixels from an image, clustering algorithms have successfully been applied to image segmentation in various fields and applications [6]. We aim to segment and classify hyper-spectral satellite images. Fully unsupervised procedures often have insufficient accurate classification results. For such a reason, a hybrid scenario between supervised and unsupervised techniques is our target where the methods applied could take into account some labels to build a classifier. We suggest a cluster-based training selection. This approach selects the training samples according to an unsupervised analysis of the data (mode seek clustering). The selected data (centers of the clusters) are likely to well represent those samples that were clustered together. This scheme was presented in [7] where a  $KNN1$  classifier was used.

Here we also introduce label propagation to adapt the method to other classifiers. For the sake of using a SVM classifier, the unlabeled data contained in each cluster is modeled regarding the distribution of their distances to their corresponding centers. The label of the center is propagated to those samples that fit this model. Besides we also test the result of assigning labels to unlabeled samples according to the result given by the cluster itself and the labels provided by the expert for the modes of clusters. For all cases, the suggested scheme is compared with the supervised state of the art classification, resulting in outperforming previous works.

A review of the sample selection scheme with its spatial improvement is presented in Section 2. Several classification alternatives are presented in Section 3. Results will be shown and analyzed in Section 5. Finally, Section 6 presents some conclusions.

## 2 Preliminaries

Nowadays, due to the improvement in the sensors, databases used for segmentation and classification of hyper-spectral satellite images are highly reliable in terms of spectral and spatial resolution. Therefore, we can consider that our feature space representation of the data is also highly reliable. On the other hand, in segmentation and classification of this kind of images the training data used has not been a concern so far, without worrying about providing the most reliable information [5]. The scheme suggested in [7] was a first attempt in this sense. It was proposed an unsupervised selection of the training samples based on the analysis of the feature space to provide a representative set of labeled data. It proceeds as follows:

1. In order to reduce the dimensionality of the problem, a set of spectral bands, given a desired number, is selected by using a band selection method. The WaLuMi band

selection method [8] was used in this case, although any other similar method could be used.

2. A clustering process is used to select the most representative samples in the image. In this case, we have used the Mode Seek clustering procedure which is applied over the reduced feature space. An improvement in the clustering process is included by adding the spatial coordinates of each pixel in the image as additional features. Since the clustering is based on distances, spatial coordinates should also be taken into account assuming the class connection principle.
3. The modes (centers of the clusters) resulting of the previous step define the training set for the next step. The expert is involved at this point, only once, by providing the corresponding labels of the selected samples.
4. The classification of the rest of non-selected samples is performed, using the training set defined above to build the classifier. Three different classification experiments have been performed here: a  $KNN$  classifier with  $k = 1$ , a direct classification with the results of the clustering process, and an extension will be presented for the use of SVM.

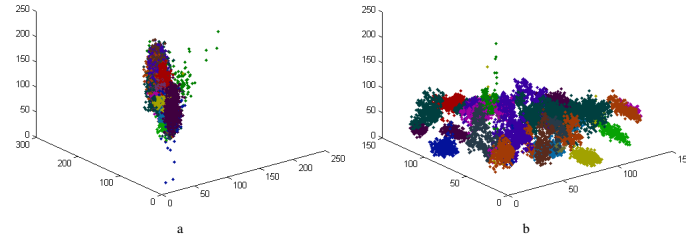
## 2.1 Mode Seek Clustering

Given a hyper-spectral image, all pixels can be considered as samples which are characterized by their corresponding feature vectors (spectral curve). The set of features defined is called the feature space and samples (pixels) are represented as points in that multi-dimensional space. A clustering method groups similar objects (samples) in sets that are called clusters. The similarity measure between samples is defined by the cluster algorithm used. A crucial problem lies in finding a good distance measure between the objects represented by these feature vectors. Many clustering algorithms are well known. A  $KNN$  mode seeking method will be used in this paper [9]. It selects a number of modes which is controlled by the neighborhood parameter ( $s$ ). For each class object  $x_j$ , the method seeks the dissimilarity to its  $s^{th}$  neighbors. Then, for the  $s$  neighbors of  $x_j$ , the dissimilarities to their  $s^{th}$  neighbors are also computed. If the dissimilarity of  $x_j$  to its  $s^{th}$  neighbor is minimum compared to those of its  $s$  neighbors, it is selected as prototype. Note that the parameter  $s$  only influences the scheme in a way that the bigger it is the less clusters the method will get since more samples will be grouped in the same cluster, that is, less modes will be selected as a result. For further information about the mode seek clustering method see [9] and [5].

## 2.2 Spatial Improvement

The clustering algorithm searches for local density maxima where the density function has been calculated using the distances for each sample in its  $s$  neighborhood using a dissimilarity measure as the distance between pairs of samples. In that difference, all features (dimensions) are considered. When features do not include any spatial information the class connection principle is missed (pixels that lie near in the image are likely to belong to the same class). Therefore, we suggest to include the spatial coordinates among the feature of the samples. See Fig 1(a) where all samples have been represented in the three first features space and in different color per class. Notice that, when

184 O. Rajadell and P. García-Sevilla



**Fig. 1.** Effects of including spatial information in the feature space. Plots show the samples of the database in the feature space, colored per class according to the ground-truth. (a) no spatial information is available. (b) spatial coordinates are included.

no spatial data is considered and all classes are located in the same space and when no prior knowledge is available for the clustering process, finding representatives for each class would be difficult since the classes themselves may lie together. Moreover, different areas of the same class may be within the same cloud. However, when spatial data is included, Fig. 1(b), the single cloud of samples is broken according to spatial distances and classes (fields) are more separable. In this sense also samples belonging to the same class but lying in different places of the image are separable.

In [7] it was suggested to weigh the spatial coordinates by an arbitrary number to reinforce two samples that are close spatially to have a closer distance and the way round. Such a weight should be decided in terms of the range of the features provided by the spectrometer so the coordinates are overweighed but they do not cause the rest of features be dismissed in the global measure.

### 3 Classification Alternatives

The whole dataset was first reduced to 10 bands using the band selection method named in Section 2. This method is used for minimizing the correlation between features but maximizing the amount of information provided, all that without changing the feature space. Clustering was carried out tuning the parameter  $s$  to get a prefixed number of selected samples. Three different classification alternatives have been used.

#### 3.1 Straightforward Schemes

1. First a  $KNN$  with  $k=1$  classification has been performed with the labeled samples as training set. This is not an arbitrary choice, because the clustering procedure used is based on densities calculated on a dissimilarity space, and therefore, the local maxima correspond to samples which minimize its dissimilarity with a high amount of samples around it. Thus, the selected samples are highly representative in distance-based classifiers.
2. Second, another classification process has been performed using the straightforward result of the clustering procedure. The expert labels the selected samples.

Then, all samples belonging to the cluster that each labeled sample is representing are automatically labeled in the same class. This provides a very fast pixel classification scheme as the clustering result is already available.

### 3.2 Extension to SVM

The scheme, as it has been presented, is not useful for classifiers that are not based on distances. However, we would like to check if providing relevant training data may be also useful for other classifiers. In this case, we extend the proposed method for SVM. For such a classifier, it would be useful to model the data shape and not their centers. Nevertheless, we do not want to increase the amount of labeled data. According to these criteria we suggest using the label of the centers as in the previous cases and using a label propagation technique to those samples fitting certain model with the aim of modeling the shape of the data and provide the SVM with a useful training set. The main idea behind label propagation is the cluster assumption. Two samples  $x_i$  and  $x_j$  have a high probability of sharing the same label  $y$  if there is a path between them in  $X$  which moves through regions of significant density [10]. Many graph-based techniques can be found in literature [11]. To propagate labels using the cluster analysis already performed and according to the main idea of label propagation, we suggest propagating the label of all cluster centers as follows:

Given the set of clusters  $W = \{w_1, \dots, w_t\}$   
 and distances  $D_i = \{d_1, \dots, d_s\}$   
 where  $d_j = \text{distance}(\text{center}_{w_i}, x_j)$  and  $x_j \in w_i$   
 we can assign the label  $y_{w_i}$  according:  
 $(x_j, y_{w_i})$  if  $0.8 * \max(D_i) \leq d_j \leq 0.85 * \max(D_i)$

We considered the possibility of propagating the label to the whole cluster or all the data included in the sphere created taking as a limit  $0.8 * \max(D_i)$ . There are two reasons for discarding these options. In the case first, propagating the label of the center to all data points in the cluster increases the errors introduced by label propagation since the further a data point is from its center the more possibilities that they do not share the same label, according the cluster assumption. As for both cases, we aimed to use a SVM as classifier and training is the most expensive step. Increasing considerably the training data has an undesired effect on the computation time. This is rather an arbitrary choice and we are currently working on the direction of how to better determine this parameter.

## 4 Data Sets

A well-known data set has been used in the experiments (see Fig 4). Hyper-spectral image 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored (the ones that cover the region of water absorption or with low SNR) [12]. The image has a spatial dimension of  $145 \times 145$  pixels. Spatial resolution is 20m per pixel. Classes range from

186 O. Rajadell and P. García-Sevilla

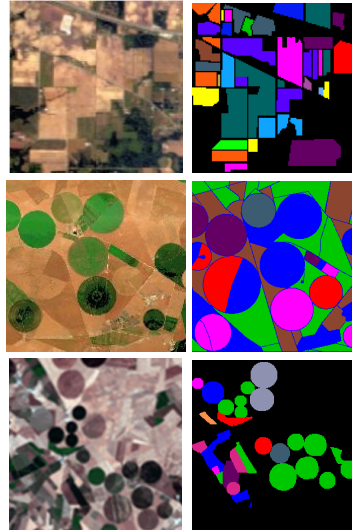
20 to 2468 pixels in size. In it, three different growing states of soya can be found, together with other three different growing states of corn. Woods, pasture and trees are the bigger classes in terms of number of samples (pixels). Smaller classes are steel towers, hay-windrowed, alfalfa, drives, oats, grass and wheat. In total, the dataset has 16 labeled classes and unlabeled part which is known as the background. This so called background will be here considered as the 17 class for the segmentation experiments.

We will analyze the details and performance for AVIRIS data set since it is a widely used data set. However we will show results with two other data sets, HYMAP and also CHRISPROBA (see Fig 4).

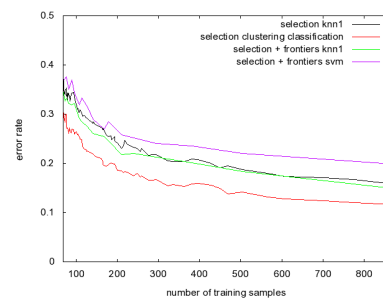
The DAISEX99 project provides useful aerial images about the study of the variability in the reflectance of different natural surfaces. This source of data, which is referred to as HyMap, corresponds to a spectral image ( $700 \times 670$  pixels and seven classes that are composed of crops and an unknown class) acquired with the 128-band HyMap spectrometer during the DAISEX99 campaign (<http://io.uv.es/projects/daisex/>). The last data set was acquired by the satellite PROBA which has a positional spectroradiometric system (CHRIS) that measures the spectral radiance, i.e., the amount of light that passes through or is emitted from a particular area. System CHRISPROBA is able to operate in several acquisition modes. The image used in this paper come from the mode that operates on an area of  $15 \times 15$  km, with a spatial resolution of 34 m, obtaining a set of 62 spectral bands that range from 400 to 1050 nm ( $641 \times 617$  pixels and nine classes that are composed of crops and an unknown class). The camera has a spectral resolution of 10 nm. Concretely, this image covering the area that is known as Barrax (Albacete, Spain) has 52 bands.

## 5 Experimental Results

In this section we will analyze the details of the method for AVIRIS data set. Later results will be shown for the other two data sets. In Fig 3 the results obtained using several classification strategies are compared:  $KNN$  using only the center of the clusters for the training set, SVM after label propagation,  $KNN$  using the same training set used for the SVM, and the classification using the plain output of the mode seek clustering. It was already shown in [7] that the scheme used with  $KNN$  clearly outperformed the random selection. Now, the classification result for the  $KNN$  classifier adding more samples in the clusters assuming the same label is very similar to the ones obtained with the  $KNN$  classifier using only the cluster centers. The SVM classifier provided the worst results in all experiments. This may be due to the fact that the double threshold scheme proposed assumes a spherical distribution of the samples around the cluster centers. However, this is not the case in general, and that is the reason why SVM cannot properly model the borders of the classes using these training samples. On the other hand, the mode seek clustering classification outperformed all other methods. The reason is that this sort of clustering is not based on the distance to a central sample in the cluster but to the distance to other samples in the clusters. When the distance to a central point is considered, a spheric distribution of the pixels around this point is assumed. However, the mode seek clustering provides clusters that may adapt to different shapes, depending on the distribution of the samples in the feature space, and these clusters can be modeled using just one sample.



**Fig. 2.** AVIRIS, HYMAP and CHRIS-PROBA data sets (respectively per row). Color composition and ground-truth.

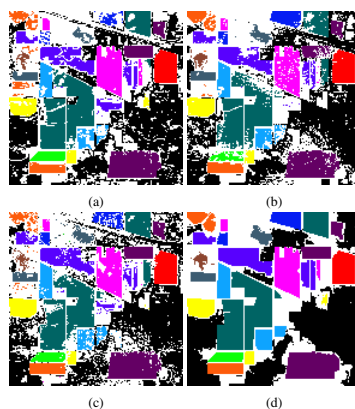


**Fig. 3.** Learning curve in terms of error rate when increasing the size of training data in number of samples selected by the scheme suggested. Different classification methods tested using the 92AV3C database.

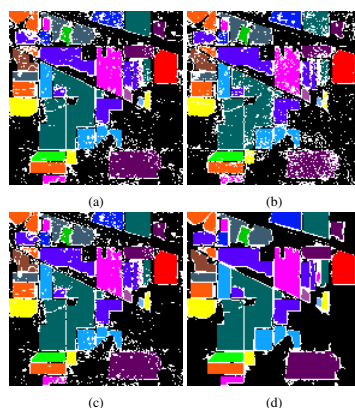
The database has 21025 samples. Fig. 4 show the classification results of several classifiers when 0.33% of the pixels in the image (69 pixels) was labeled by the expert. The classification errors are shown as white pixels. It can be noted that the clustering



188 O. Rajadell and P. García-Sevilla

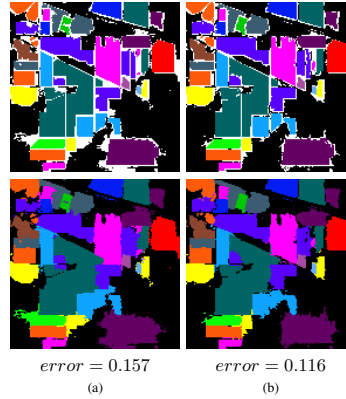


**Fig. 4.** Segmentation-classification results using 0.33% of data for the selected training set using several classifiers. (a) *KNN* using the cluster centers. (b) *SVM* (c) *KNN* using the same training set as for the *SVM* (d) mode seek clustering.



**Fig. 5.** Segmentation-classification results using 4% of data for the selected training set using several classifiers. (a) *KNN* using the cluster centers. (b) *SVM* (c) *KNN* using the same training set as for the *SVM* (d) mode seek clustering.

classifier outperformed the other classifiers not only in the percentage of classification rate but also providing smooth compact regions in the image. Similar results can be seen in Fig. 5 where 4% of the pixels in the image was labeled, where the classification



**Fig. 6.** Segmentation-classification results using different amounts of data for the selected training set using the proposed scheme and the clustering based classification. (a) Using 2% of the data. (b) Using 4% of the data.

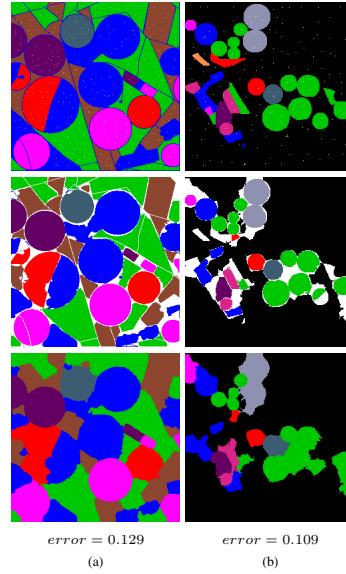
errors tend to concentrate in the borders of the different regions in the image. Note that the segmentation results are quite smooth even for the background class.

Let's consider the 2% of the samples and the cluster-based classification. See results in Fig 6(a). Observe the top left part of the image where the selection manages to detect all of them although the classes are lying one next to each other and their size is not big. The best result is presented in Fig 6(b), it is the classification-segmentation result for the 17-classes problem using 4% of the data. The overall error rate is 0.116 and the most relevant error is the lost of very small classes that cannot be found by the clustering. In Table 1 the results per class are presented for different sizes of the training set using cluster classification. Observe that the accuracy per class of a reduced training set is good when the class has been detected by the cluster. As long as one class is missed in the selection of the training data, this class will be entirely misclassified.

A brief overview of the results for the other two data sets can be found in Fig 7. This data sets have higher spatial resolution and better results were expected for them. Indeed, error rates of 0.1 are reached for both when less than 0.5% of the data is used for training. In this cases, all classes are big enough in number of samples and there are no classes missed in the selection process. Again, errors are placed at the borders of the areas. Note that in HYMAP data set there is an area defined in the groundtruth that draws a line around all visible shapes and it is labeled. This area is too narrow and always confused with the adjacent classes, for such an example of class distribution this method will have difficulties since their samples are spatially very close to other areas and they never form a structure big enough to be detected by itself.

In Table 1 where the error rate per class is shown, we can see that the results obtained using 2% of the samples are already comparable in terms of per class accuracy with

190 O. Rajadell and P. García-Sevilla



**Fig. 7.** Segmentation-classification results for other data sets selecting the training set using the proposed scheme and the clustering based classification. The training set selected is shown at the first row, at the second row the error resulting is presented in white and last row shows the classification result for (a) Using 0.312% of the data set HYMAP. (b) Using 0.244% of the data set CHRIS-PROBA.

results obtained in supervised scenarios using 5% of the data [11]. Notice that classes with only one spatial area are well classified with few samples needed, such as Alfalfa, Wheat, Hay-windrowed, Grass/pasture-mowed and Corn. Some of them (as Wheat and Hay-windrowed) were already well classified when only 0.33% training data was used. The rest of the classes are divided in different spatial areas and their detection is highly dependant on the size of the area and the amount of different classes that surrounds them. Soybeans-min-till class is from the beginning well classified with only 10 samples, this is a large class whose different areas in the image are also large and well defined. The same can be concluded for other classes like Bldg-Grass-Tree-Drives or Woods. However, class Soybeans-clean till is confused with the classes around since the areas where it lies in are small despite of being a big class. The background is a special case, although it is treated here as a single class for segmentation purposes, it consists of different areas with probably considerably different spectral signatures and, if a part of it would be missing in the training data, that part will be misclassified.

**Table 1.** Accuracy per class for the 17 classes classification of the AVIRIS dataset using 12 features (ten spectral features and two spatial coordinates). For a training sets of 0.33%, 2% and 4% of the data using the clustering-based classifier.

classes	0.33% of training data		2% of training data		4% of training data	
	training/total	error	training/total	error	training/total	error
Heterogenous background	22/10659	0.432	171/10659	0.262	367/10659	0.193
Stone-steel towers	0/95	1	2/95	0.139	5/95	0.033
Hay-windrowed	2/489	0.004	10/489	0.004	25/489	0.004
Corn-min till	5/834	0.214	18/834	0.076	40/834	0.045
Soybeans-no till	5/968	0.185	25/968	0.060	40/968	0.072
Alfalfa	0/54	1	1/54	0.038	3/54	0.039
Soybeans-clean till	2/614	0.488	15/614	0.066	28/614	0.056
Grass/pasture	3/497	0.105	12/497	0.064	28/497	0.042
Woods	6/1294	0.023	29/1294	0.034	58/1294	0.026
Bldg-Grass-Tree-Drives	3/380	0.021	9/380	0.011	12/380	0.011
Grass/pasture-mowed	0/26	1	1/26	0.040	1/26	0.040
Corn	1/234	0.601	6/234	0.070	10/234	0.049
Oats	0/20	1	0/20	1	0/20	1
Corn-no till	6/1434	0.278	35/1434	0.067	63/1434	0.035
Soybeans-min till	10/2468	0.069	70/2468	0.023	143/2468	0.018
Grass/trees	4/747	0.067	18/747	0.033	34/747	0.042
Wheat	1/212	0.009	7/212	0.005	11/212	0.005
Overall error		0.299		0.156		0.116

## 6 Conclusions

A training data selection method has been proposed in a segmentation classification scheme for scenarios in which no prior knowledge is available. This aims at improving classification and reducing the interaction with the expert who would label a very small set of points only once. This is highly interesting when expert collaboration is expensive. To get representative training data, mode seek clustering is preformed. This type of clustering provides modes (representative samples) for each cluster found in the feature space and those modes are the selected samples for labeling. Thanks to a spatial improvement in the clustering, the modes provided do not contain redundant training information and can represent different spatial areas in the image that belong to the same class. The training selection has been used over several classifiers. We have experimentally proved that distance based classifiers are more adequate than SVM for such an approach. Furthermore, we have also shown that the classification obtained from the mode seek clustering outperformed the simple distance based classifiers because it better adapts to the shapes of the clusters in the feature space.

All classification strategies benefit from the selection of the labeled data to improve their performances. They provide very good results even with less labeled data than provided in other scenarios where training data was randomly selected.

192 O. Rajadell and P. García-Sevilla

## References

1. Tarabalka, Y., Chanussot, J., Benediktsson, J.A.: Segmentation and classification of hyperspectral images using watershed transformation. *Patt. Recogn.* 43, 2367–2379 (2010)
2. Plaza, A., et al.: Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment* 113, 110–122 (2009)
3. Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., Emery, W.: Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 47, 2218–2232 (2009)
4. Li, J., Bioucas-Dias, J., Plaza, A.: Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE TGRS* 48, 4085–4098 (2010)
5. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 603–619 (2002)
6. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 898–916 (2011)
7. Rajadell, O., Dinh, V.C., Duin, R.P., García-Sevilla, P.: Semi-supervised hyperspectral pixel classification using interactive labeling. In: *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS 2011* (2011)
8. Martínez-Usó, A., Pla, F., Sotoca, J., García-Sevilla, P.: Clustering-based hyperspectral band selection using information measures. *IEEE Trans. on Geoscience & Remote Sensing* 45, 4158–4171 (2007)
9. Cheng, Y.: Mean shift, mode seek, and clustering. *IEEE Transaction on Pattern Analysis and Machine* 17, 790–799 (1995)
10. Chapelle, O., Schölkopf, B., Zien, A.: *Semi-Supervised Learning*. MIT Press, Cambridge (2006)
11. Chi, M., Yu, X.H.S.: Mixture model label propagation. In: *19th ACM International Conference on Information and Knowledge Management*, pp. 1889–1892 (2010)
12. Landgrebe, D.A.: *Signal Theory Methods in Multispectral Remote Sensing*, 1st edn. Wiley, Hoboken (2003)

## Training Data Selection for Cancer Detection in Multispectral Endoscopy Images

Cuong V. Dinh<sup>1,2</sup>, Marco Loog<sup>1</sup>, Raimund Leitner<sup>2</sup>, Olga Rajadell<sup>3</sup>, Robert P.W. Duin<sup>1</sup>

<sup>1</sup>Pattern Recognition Laboratory, Delft University of Technology, the Netherlands

<sup>2</sup>Carinthian Tech Research AG, Austria

<sup>3</sup>Institute of New Imaging Technologies, University Jaume I, Spain

### Abstract

*Multispectral endoscopy images provide potential for early stage cancer detection. This paper considers this relatively novel imaging technique and presents a supervised method for cancer detection using such multispectral data. The data under consideration include different types of cancer. This poses a challenge for the detection as different cancer types may exhibit different spectral signatures. Consequently, it is not always feasible to transfer the knowledge learnt from one data set to another data set. In our approach, we select suitable training data for a given test set based on a similarity measurement between data sets. Experimental results demonstrate that the classification results can be significantly improved if a few data sets that are presumably similar to a given test set are selected for training instead of using all available data sets.*

### 1 Introduction

Early cancer detection plays an important role in increasing the chance for successful cancer treatment. A common technique for early cancer diagnosis is taking biopsies, which requires physical removal of specimens followed by a histopathological analysis [6]. It is difficult to determine the dysplastic and malignant regions for biopsies and therefore the procedure may have to be repeated many times, which delays the necessary treatment.

Optical techniques, such as the autofluorescence spectroscopy, have been investigated for early cancer diagnosis. Autofluorescence is the light emission of specific substances of biological tissues, e.g. porphyrins and proteins if the tissues are excited by a light source. Those substances then emit light of specific wavelengths. The spectra of the tissues then correspond

to different wavelengths measured by the spectroscopy. Previous studies, e.g. [2] have shown that there is a significant difference in the fluorescent properties, such as their spectral shape and intensity, between malignant and normal tissues. Therefore, they have been used to identify early instances of diseases in the colon, larynx, lung, and other organs.

The advantage of optical techniques lies in their potential to perform *in vivo* detection without the need for tissue removal. Therefore, they facilitate the determination of the dysplastic and malignant regions for the biopsy. These spectroscopic diagnosis techniques are often referred to as point-measurement methods as they attempt to obtain the spectra of a single tissue.

Multispectral/hyperspectral endoscopy techniques developed recently provide three-dimensional images of the area of interest in both spatial and spectral domains [3, 4, 6]. Multispectral images provide richer information than point-measurement techniques as they are able to acquire the spectra of thousands to millions of malignant and normal pixels at the same time. In [4], a thresholding algorithm is used to assign pixels to normal/malignant spectra based on the observation that the intensity of a malignant area is brighter than that of a normal area. In this paper, we present a supervised method, in particular, we focus on the issue of transferring knowledge among data sets.

The data under consideration consist of eight multispectral endoscopy images belonging to different types of cancers. As different cancer types may exhibit different spectral signatures [1], the discriminant information between normal and malignant tissues learnt from a data set may not be applicable to another data set. We address this problem by selecting suitable training data sets for a given test set. Data sets are only selected for training if they are similar to the test set, i.e. they stay close to it in the feature space. Experimental results show that the classifications can be significantly improved if a few data sets which are similar to a test set

are selected for training instead of using all data sets.

## 2 Materials

Data were collected from patients with different kinds of cancer at the hospitals for otolaryngoscopic and thorax surgeries in Stuttgart, Germany. Multispectral images of the investigated tissue areas were recorded quickly after surgery so the *in vivo* conditions of the tissues are commonly believed to be conserved. The hyperspectral images were produced using an electron multiplying charge coupled device (EMCCD) camera, with a resolution of  $1002 \times 1004$  pixels, an acousto-optic tunable filter (AOTF) with wavelengths ranging from 400nm to 650nm (FWHM 5nm), and a 10 mm laparoscope with a 300W Xenon light source.

The eight data sets under consideration (called M1, M2  $\dots$  M8) belong to different types of cancer: Laryngeal cancer (data sets M3, M4, M5, and M8), Pharyngeal cancer (M1), Esophageal cancer (M2), Diaphragm cancer (M6), and Parotid cancer (M7). For the M4 data set, the exact boundary of the cancer area is unclear since the cancer tissue is under the surface. Therefore, it is not easily detectable by a non-penetrating optical method. All data sets are acquired in a white light condition and the number of spectral bands is 51.

## 3 Methods

### 3.1 Data preprocessing

First, each reflectance spectrum is normalized using the area under the curve normalization in the spectral domain. Second, spectra corresponding to the specular reflection are removed by a simple thresholding algorithm. Third, the principal component analysis (PCA) is used to reduce the number of features from the original space. The reconstruction of all data sets are then based on their first eight eigenvectors which preserve 99% of the total variance. Finally, a unit variance normalization is applied to each data set so that each spectral band has a zero mean and a unit variance. The main aim of this normalization is to align all the data sets, i.e. to force them to stay close to each other in the feature space.

### 3.2 Data selection for training

As the data sets are different from one to another with respect to their class distributions, the discriminant information between normal and malignant tissues learnt from a data set might not be suitable for another data set. Therefore, it is essential to select suitable training sets for a given test set. We first use the Gaussian

data domain description [5] to model the distribution of each data set. Denote  $q$  the percentage of outliers in each data set, a pixel is considered as an outlier if its probability density  $p(x_i)$  is smaller than a threshold  $\theta$  determined by:

$$\frac{1}{N} \sum_{i=1}^N h(\theta - p(x_i)) = q$$

where  $N$  is the total number of pixels in the data,  $h(\cdot)$  the unit step function, and  $p(x_i)$  the probability density of pixel  $x_i$ . We then measure the similarity between two data sets by the fraction of pixels they share in their data domain. For two data sets  $M_i$  and  $M_j$ , we calculate  $M_{ij}$  the set of all pixels in  $M_i$  that belong to the domain of  $M_j$  and  $M_{ji}$  the set of all pixels in  $M_j$  that belong to the domain of  $M_i$ . The similarity between  $M_i$  and  $M_j$  denoted by  $S_{ij}$  is defined as:  $S_{ij} = |M_{ij}|/|M_i| + |M_{ji}|/|M_j|$ . The similarity among data sets are then used as the criterion to select the training set for a given test set. Note that we model a data set using all the pixels contained. It is therefore possible to measure the similarity between any two data sets, e.g. between a training set and a test set, even when we do not have label information of the test set.

## 4 Experimental results

As knowledge about the prior probabilities of normal and malignant classes is not available, we set them to be equal in all experiments. We use the quadratic discriminant classifier (QDC) for the classification between normal/malignant tissues.

### 4.1 All available data sets are used for training

We first evaluate the classification results for two training scenarios: i) training and test data are from the same data set, i.e. a part of a data set is used for training and the remainder is for testing; ii) training and test data are from different data sets. For the latter, we follow the leave-one-dataset-out cross validation configuration, i.e. seven data sets are used for training and the remaining data set is used for testing. Moreover, for the second scenario we also investigate the influence of the unit variance normalization in the data preprocessing step. Since the QDC is invariant to affine transformations, the classification results for the first scenario remain unchanged whether this normalization is applied or not. Table 1 shows the classification results with respect to different training and normalization options. The table clearly shows the difference between

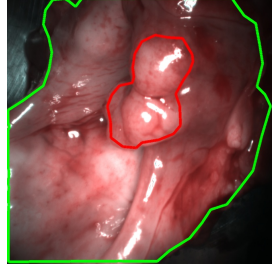


Figure 1. Reconstructed color image of the data set M8. Normal and malignant areas are marked by green and red contours.

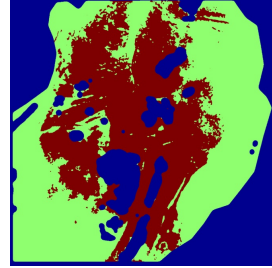


Figure 2. Classification result for the M8 data set using the unit variance normalization and QDC.

Table 1. Error rate (%) for different training and normalization options

Training scenario	Normalization	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Same set	No	09.9	11.0	16.0	10.1	05.8	15.8	10.0	07.6	10.8
Different sets	No	39.8	48.9	34.0	28.8	51.6	46.0	30.2	22.8	37.7
Different sets	Yes	<b>30.1</b>	<b>26.5</b>	<b>36.0</b>	<b>29.6</b>	<b>17.6</b>	<b>38.1</b>	<b>30.5</b>	<b>23.3</b>	<b>29.0</b>

Table 2. Error rate (%) when training data selection is used

	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Case 1	42.0	27.0	45.4	51.2	27.2	42.7	31.9	24.0	36.4
<b>Case 2</b>	<b>26.8</b>	25.4	<b>30.1</b>	<b>24.1</b>	26.5	50.9	<b>26.5</b>	<b>16.0</b>	<b>28.3</b>
Case 3	30.8	27.6	32.6	28.4	20.1	45.6	28.9	18.4	29.1
Case 4	31.8	26.2	39.0	25.2	18.7	44.0	29.8	24.3	29.9
Case 5	26.8	<b>25.3</b>	36.9	27.7	19.0	39.6	34.3	25.1	29.3
Case 6	29.2	24.7	36.0	28.7	<b>17.2</b>	<b>36.1</b>	30.9	24.7	28.4
Case 7	30.1	26.5	36.0	29.6	17.6	38.1	30.5	23.3	29.0

Table 3. Best error rate (%) and the corresponding number of data sets used for training

	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Errors	21.6	20.0	28.2	22.1	13.3	29.7	23.8	13.8	21.5
#training sets	3	2	2	2	2	4	3	2	2.5

the two training scenarios. The error rate increases substantially when training and test data are not from the same data set as they are far different from each other. In addition, the unit variance normalization is demonstrated to significantly improve the classification results when the training and test data are from different data sets. Therefore, we apply the normalization step in all of the following experiments. Note that we also used other classification methods, such as Parzen classifier and the linear SVM; however, they often perform worse

than the QDC. Figure 1 displays the reconstructed color image for the data set M8 and its normal and malignant areas. The border of the malignant area has been annotated by a medical expert. The classification on this data set is shown in Figure 2. Detected malignant and normal areas are displayed in red and green, respectively. Blue depicts background. The blue areas within the tissues correspond to the specular reflections. They are removed during the preprocessing step mentioned in Section 3.1 and therefore not considered in the classifica-



tion. The figure shows that the detected malignant area tends to expand into the normal area. One reason might be that malignant tissues in the training data are different from one another as they belong to different cancer types. Thus, they exhibit mixture data distribution. However, the QDC assumes only a single Gaussian for each malignant/normal class. Consequently, the estimated distribution of the malignant class becomes more flat and therefore more tissues become false positive.

#### 4.2 Training data selection

We evaluate the classification results when the training data contain similar data sets for a given test data set. We model the data sets by using the Gaussian domain description in which the percentage of outlier  $q$  is set to 0.1. For each data set, we first selected the training data as the most one, two  $\cdots$  seven similar data sets (denoted by Case 1, 2  $\cdots$  7) according to the similarity measurement defined in the Section 3.2. The QDC is then trained on the selected training data and subsequently used for the classification of normal/malignant tissues for the data set under consideration. Table 2 shows the error rates for all seven cases. Numbers in bold emphasize the best results achieved for each data set in all cases. Note that Case 7 corresponds to the results shown in the third row of Table 1 as all seven data sets are included in the training data. On average, the best classification results are obtained if the two most similar data sets are used for training (Case 2). Increasing the number of training data sets then, in most of the time, worsens the classification as irrelevant data are included in the training process. Case 2 yields the best results for five over eight data sets. Case 2 does not perform well on the data set M6 as the data set itself is challenging: the cancer type (diaphragm) is totally different from the other cancer types.

We also carry out experiments in which for each data set, the training data is manually selected according to the classification results. Table 3 shows the best error rates and the corresponding number of data sets used for training. Similar to the above results (cf. Table 2), the classifier performs best when two or three data sets are selected for training. We also noted that for any of the three data sets M3, M5, and M8, the best classification result is achieved if the other two data sets are included in the training data except for the M5 where the best training set contains M7 and M8 yielding an error rate of 13.3%. Nevertheless, the training set containing M3 and M8 produces a comparable result of 15.1% error rate. This confirms the fact that the three data sets are similar as they exhibit the same type of cancer (Laryngeal cancer).

## 5 Conclusion

This paper presents a study of normal/malignant tissue classification for eight multispectral endoscopy data sets in a supervised manner. The data are heterogeneous as they are collected from different patients and with different types of cancer. We showed that the classification result is improved if a subset of the data that are similar to the test set is used for training (cf. Table 2 & 3). In other words, it is not always good to combine all available data for training as the difference between the data sets may result in poor classification.

We introduce an approach to select training data based on the similarity between data sets using the Gaussian data domain description. Experimental results show that the method substantially improves the classification results for our heterogeneous data. Note that we measure the similarity between data sets based on all the pixels, i.e. from both normal and malignant classes. For data from a patient who does not have cancer, all the pixels should fall into the normal region of the selected training data; therefore, our method correctly classifies the data set as normal.

In the present paper we use PCA to reduce the dimensionality of the feature space. To find subspaces that provide discriminant information between normal/malignant tissues in the data may also improve the performance of the classifiers. Finally, more data sets are essential to fully evaluate the applicability of our method.

## References

- [1] S. Demos, R. Gandour-Edwards, R. Ramsamooj, and R. deVere White. Near-infrared autofluorescence imaging for detection of cancer. *Journal of biomedical optics*, 9:587, 2004.
- [2] M. Harries, S. Lam, C. MacAulay, J. Qu, and B. Palcic. Diagnostic imaging of the larynx: autofluorescence of laryngeal tumours using the helium-cadmium laser. *The Journal of Laryngology & Otology*, 109(02):108–110, 1995.
- [3] R. Leitner, T. Arnold, and M. De Biasio. High-sensitivity hyperspectral imager for biomedical video diagnostic applications. In *Proceedings of SPIE*, volume 7674, 2010.
- [4] M. Martin et al. Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection. *Annals of biomedical engineering*, 34(6):1061–1068, 2006.
- [5] D. Tax. *One-class classification; Concept-learning in the absence of counter-examples (Chapter 3)*. PhD thesis, Delft University of Technology, 2001.
- [6] T. Vo-Dinh. A hyperspectral imaging system for in vivo optical diagnostics. *Engineering in Medicine and Biology Magazine, IEEE*, 23(5):40–49, 2004.

